Universität Hildesheim
Fachbereich III: Informations- und Kommunikationswissenschaften

Magisterarbeit
Internationales Informationsmanagement

Einsatz von Entry Vocabulary Modulen zur Optimierung von mehrsprachigen Information Retrieval Systemen am Beispiel der Datenbank des Fachinformationsverbundes Internationale Beziehungen und Länderkunde

Benjamin Berghaus

Hildesheim, September 2006

Erstgutachter: Dr. habil. Thomas Mandl
Zweitgutachterin: Prof. Dr. Christa Womser-Hacker

Inhaltsverzeichnis

Abbildungsverzeichnis

Hinweis: Alle Abbildung mit Ausnahme von Abbildung 2.1 sind eigene Darstellungen. Abbildung 1.1 wurde aus den Logos der Kooperationspartner im Fachinformationsverbund zusammengestellt.

Tabellenverzeichnis

Hinweis: Die Inhalte aller abgebildeten Tabellen, bis auf Tabelle 2.1, sind Ergebnisse eigener Untersuchungen oder der Entwicklungsarbeit.

Teil I.

Einleitung

Vorwort und Danksagung

Durch die interdisziplinäre Ausbildung und den besonderen Fokus auf der Anwendung von Informationswissenschaften im Fach Internationales Informationsmanagement erhalten die Studenten dieses Studienfeldes einen fundierten Einblick in und ein Gespür für die Verbindung des wissenschaftlichen Umgangs mit Informationen mit der praktischen Applikation des Gelernten. Die vorliegende Arbeit versucht diese beiden Aspekte zu verbinden. Sie wäre nicht zustande gekommen, hätte es nicht aus beiden Bereichen tatkräftige Unterstützung gegeben.

So möchte ich mich auf der Seite der praktischen Anwendung herzlich für die Zusage zur Kooperation, die Bereitstellung der im Folgenden verwendeten Datengrundlage, die Übernahme des Großteils der Evaluierungsleistung und die stets freundliche, zuvorkommende Unterstützung bei Herrn Michael Kluck, Frau Dr. Petra Galle und ihren Kollegen bei der Stiftung Wissenschaft und Politik (SWP) in Berlin bedanken. Ohne die umfassende und engagierte Unterstützung dieser Arbeit durch den Fachinformationsbereich der Stiftung Wissenschaft und Politik wäre eine Forschungsarbeit im Überschneidungsbereich von Politikwissenschaften und Informationswissenschaften nicht möglich gewesen. Darüber hinaus möchte ich mich für die freundliche Unterstützung im Rahmen der Vorbereitung der Relevanzbewertung durch Herrn Stefan Bärisch am Informationszentrum Sozialwissenschaften in Bonn bedanken.

Auf der anderen, wissenschaftlichen Seite möchte ich ebenso herzlich Danksagen für die Zusage zur Betreuung und die umfassende Unterstützung durch meine beiden Prüfer, Herrn Dr. habil. Thomas Mandl und Frau Prof. Dr. Christa Womser-Hacker an der Universität Hildesheim. Auf der Basis eines Themenvorschlags durch Herrn Mandl wurde diese Arbeit konzipiert und durch ihn wurde der Kontakt zur SWP hergestellt. Ohne die äußerst zahl- und hilfreichen Beratungsgespräche und die wertvollen Hinweise auf Ansätze für weitere Verbesserungsmöglichkeiten des Systems hätte ich diese Arbeit weder beginnen noch abschließen können.

Über diese Arbeit

Zusammenfassung

Diese Arbeit befasst sich mit der Entwicklung und Evaluierung eines Information Retrieval Systems. Im Vordergrund steht hierbei die Entwicklung eines dynamischen Entry Vocabulary Moduls (EVM). Im Laufe der Arbeit soll ein durch ein EVM unterstütztes Retrieval System in seiner Leistungsfähigkeit untersucht werden. Bei der Datenbasis, für die das System entwickelt und auf der es evaluiert wird, handelt es sich um den Datenbestand des Fachinformationsverbundes Internationale Beziehungen und Länderkunde, die als Auszug vom Kooperationspartner Stiftung Wissenschaft und Politik bereitgestellt wurde. Außerdem wurde das System auf die German Indexing and Retrieval Testdatabase (GIRT) angewendet. Diese Forschungsarbeit hat neben vielen weiteren Erkenntnissen ergeben, dass im direkten Vergleich zwischen zwei Evaluierungsläufen des Systems ohne zusätzliche Module und drei Läufen mit zusätzlichen Modulen die Retrievalleistung gesteigert werden konnte und das EVM daran maßgeblichen Anteil hatte.

Abstract

This thesis deals with the development and evaluation of an information retrieval system encorporating a dynamic entry vocabulary module (EVM). Furthermore, the evaluation of the system is covering an analysis of the potential performance gains that can be reached employing an evm. The data that the system is being developed for is the database of the Fachinformationsverbund Internationale Beziehungen und Länderkunde which has been kindly provided by the cooperating partner Stiftung Wissenschaft und Politik, Berlin. The system has also been evaluated on the data of the German Indexing and Retrieval Testdatabase (GIRT). In conclusion, this thesis compares two simpler evaluation runs without any additional modules and three advanced evaluation runs with all additional modules enabled. The evaluation resulted, alongside a series of other results and findings, in an advantage for the advanced runs and hence supports the theory of the potential of an entry vocabulary module.

Zielsetzung

Die Zielsetzung dieser Arbeit teilt sich in zwei Aufgabenkomplexe:

Der praktische Aspekt der Arbeit umfasst, ein Retrievalsystem auf Basis der Java-Klassenbibliothek Lucene zu entwickeln. Das Retrievalsystem wird für und auf der Basis der Datengrundlage des Fachinformationsverbundes für Internationale Beziehungen und Länderkunde entwickelt. Für dieses Retrievalsystem soll im Speziellen ein Entry Vocabulary Modul entwickelt werden. Eine Erweiterung des praktischen Kapitels der Arbeit ist die Evaluierung des Systems in mehreren Evaluierungsläufen, sowohl auf der Datengrundlage des Fachinformationsverbundes, als auch von GIRT. Die Evaluierung soll sowohl im Hinblick auf die einsprachige, als auch auf die mehrsprachige Retrievalqualität des Information Retrieval Systems eine Aussage liefern können.

Der analytische Aspekt der Arbeit umfasst die Konzeption eines geeigneten Entry Vocabulary Moduls für die vorliegende Datengrundlage, die Vorbereitung sowie die Auswertung der Evaluierung und die Ableitung von weiteren Verbesserungsmöglichkeiten im Ausblick.

Zu erforschende Fragestellungen

Die zentralen, zu erforschenden Fragestellungen für diese Arbeit lauten:

1. Welche Ansätze für den Information Retrieval Prozess lassen sich aus dem speziellen Anwendungsfall des Retrievals von Dokumenten mit inhaltlichem Bezug auf Außen- und Sicherheitspolitik ableiten?

2. Welche Retrievalstrategien lassen sich für die Entwicklung eines Retrievalsystems für Referenzdatenbanken ableiten?

3. Lässt sich die Retrievalleistung eines IR-Systems durch die Anwendung eines Entry Vocabulary Moduls verbessern?

4. Wie verändert sich die Retrievalleistung eines IR-Systems mit Entry Vocabulary Modul, wenn die Parameter des Systems geändert werden?

5. Welchen Einfluss haben die eingesetzten Module auf die mehrsprachige Retrievalleistung?

6. Welche zusätzlichen Ansätze zur Steigerung der Retrievalleistung sind während der Entwicklung des Systems deutlich geworden?

Die Klärung dieser Fragen wird in der Auswertung der Evaluierung in Kapitel 10.3.2 auf Seite 91 erfolgen.

Gliederung der Arbeit

Diese Arbeit gliedert sich in fünf Teile. Der erste Teil umfasst diese grundlegende Einleitung.

Im zweiten Teil ab Seite 2 werden zunächst Hintergrundinformationen über den Kooperationspartner Stiftung Wissenschaft und Politik und die Referenzdatenbank des Fachinformationsverbundes Internationale Beziehungen und Länderkunde gegeben. Außerdem werden in gebotener Kürze die Grundlagen des Information Retrieval und der im weiteren Verlauf der Arbeit eingesetzten Technologie des Entry Vocabulary Moduls beschrieben.

Der dritte Teil ab Seite 28 beschreibt die Entwicklung des im Rahmen dieser Arbeit programmierten Retrievalsystems ausgehend von der Analyse der zugrundeliegenden Daten und der Darstellung unterschiedlicher Nutzungssituationen von verschiedenen Nutzergruppen. Auf Basis der Analyse der vorliegenden Datengrundlage wird dann das Konzept für ein dynamisches Entry Vocabulary Modul beschrieben. Außerdem werden der Indexierungsprozess und die Architektur des Suchprozesses inklusive aller verwendeten Module dokumentiert.

Der vierte Teil ab Seite 56 beinhaltet die Dokumentation der Evaluierung des beschriebenen Systems. Neben einer Analyse der eingesetzten Evaluierungsläufe geht eine umfassende Auswertung der Daten darauf ein, ob eine Steigerung der Retrievalqualität erzielt werden konnte, welche Parameterkombinationen erfolgreich waren und welche Anfragen besonders von den zusätzlich eingesetzten Modulen profitieren konnten. Abschließend werden die Evaluierungsergebnisse zusammengefasst und Antworten auf die zentralen Fragestellungen aus Abschnitt I gegeben.

Im fünften Teil ab Seite 96 werden ausblickend diverse, bei der Entwicklung und Evaluierung aufgekommenen, Ansätze zur Verbesserung des Retrieval Systems und damit des Retrievalergebnisses dargestellt. Daraufhin beschließt das Fazit die Arbeit mit einer Zusammenfassung der Erkenntnisse und Empfehlungen für die weitere Entwicklung des Konzepts.

Im Anhang finden sich die Auflistung und Analyse der Evaluierungsanfragen sowie das Literaturverzeichnis.

Dieser Arbeit liegt eine CDROM bei, auf der sich der Quellcode der eingesetzten Klassen, die Ergebnisse und Protokolle der Evaluierungsläufe sowie die im Rahmen der Enwicklung erstellten Indices der GIRT-Daten und des Thesaurus befindet.

Teil II.

Grundlagen

1. Stiftung Wissenschaft und Politik

1.1. Entstehung und Entwicklung der Stiftung

Deutschland stellte nach dem Ende des zweiten Weltkriegs in besonderer Art und Weise die Front zwischen Ost und West dar: Der nach dem Kriegsende unmittelbaren Aufteilung Deutschlands durch die Siegermächte England, Frankreich, Russland und den Vereinigten Staaten von Amerika folgte 1948/49 zunächst die Blockade West-Berlins durch die sowjetischen Besatzer. Es kam durch die Etablierung des Ost-Magistrats zur politischen Spaltung der Stadt. Im Jahr 1961, während der dritten Berlinkrise, wurde durch den Bau der Berliner Mauer die Spaltung Deutschlands in die Deutsche Bundesrepublik und die Deutsche Demokratische Republik Realität.

> „Es waren die Teilung Deutschlands und die kommunistische Bedrohung, die den Wunsch nach politikbegleitender Forschung gefördert hatten, einen Wunsch, der sich 1962 in der Gründung der Stiftung Wissenschaft und Politik [...] niederschlug.“ (Heinrich Vogel, Mitglied des Vorstands der SWP in [26], Seite 23)

Gegenwärtig war also nicht nur die Geschichte der erst kürzlich vergangenen Weltkriege, sondern im Besonderen auch das Resultat aus den Kriegen. Der außen- und sicherheitspolitische Dipol NATO (seit 1949) und Warschauer Pakt (seit 1955) führte zu der außerordentlich konfrontativen Situation des kalten Krieges, die es zunächst zu entschärfen und schließlich politisch zu lösen galt.

In dieser Situation wurde die Stiftung Wissenschaft und Politik und mit ihr das Deutsche Institut für Internationale Politik und Sicherheit im Jahr 1962 in Ebenhausen bei München gegründet. Zu diesem Zeitpunkt war die SWP eine privat geförderte Initiative. Nur drei Jahre später, im Januar 1965, beschloss der Bundestag der Bundesrepublik Deutschland einstimmig, dieser Initiative beizutreten und so die Stiftung finanziell zu fördern. Weitere finanzielle Unterstützung erhielt das Institut aus Drittmitteln von in- und ausländischen Gesellschaften zur Forschungsförderung.

Seitdem haben sich zentrale politische Machtgefüge grundlegend verändert. Alte Konfliktherde sind erloschen und neue Spannungsfelder entstanden. Die SWP hat im Jahr 2001 ihren Sitz in die Hauptstadt verlagert und sich innerhalb der vergangenen 40 Jahre zu dem „größten außenpolitischen Forschungsinstitut [...] in ganz Westeuropa“ (Christoph Bertram, Vorstandsvorsitzender der SWP in [26], Seite 28) entwickelt und stellt für Regierung, Ministerien und den Bundestag eine wichtige, beratende Instanz in außen- und sicherheitspolitischen Fragen dar.

1.2. Heutige Situation und Aufgabenstellung der Stiftung

Heute, gut 40 Jahre nach der Gründung der Stiftung, verfügt die SWP durch Zuwendungen des Bundes und darüber hinaus durch Drittmittel über ein Budget von rund 10 Millionen Euro sowie einen nominellen Stab von rund 140 Mitarbeitern.

Die Stiftung Wissenschaft und Politik und damit das Deutsche Institut für Internationale Politik und Sicherheit hat zur Kernaufgabe, Forschung im Feld der angewandten Politikwissenschaft zu betreiben. Die SWP betreibt dabei keine Auftragsforschung: Vielmehr werden die aus den Forschungsergebnissen entwickelten Beratungsdienstleistungen durch die Kunden des Instituts nachgefragt. Die externen Auftraggeber sind vorrangig der Deutsche Bundestag, die Bundesregierung und die Bundesministerien wie das Auswärtige Amt und das Bundesministerium für Verteidigung.

> „An der Schnittstelle zwischen Wissenschaft und Politik wird die Bewahrung der geistigen Unabhängigkeit - das heißt Offenheit für alternative Interpretationen der Fakten, für mehr als eine Variante der Zukunft - zur Herausforderung.“ (Heinrich Vogel, Mitglied des Vorstands der SWP in [26], Seite 24)

Aus dieser Kernaufgabe ergibt sich die Position der SWP als wissenschaftliche Beratungsinstanz für die politischen Organe der Bundesregierung und der Verwaltung. Die SWP ist dabei keine Forschungseinrichtung, die einem speziellen Ressort der Bundesregierung oder der Verwaltung zugeordnet ist. Es handelt sich vielmehr um eine Einrichtung, die unabhängig von Weisungen ist und nicht mit der Erfüllung hoheitlicher Aufgaben betraut werden kann. So ist die Unabhängigkeit der SWP garantiert, die eine notwendige Voraussetzung für die unbeeinflusste Beratungsdienstleistung darstellt (vgl. Frank-Walter Steinmeier in [26], Seite 18). Der wissenschaftliche Anspruch der Stiftung führt darüber hinaus zu einer klaren methodischen und zielorientierten

Unterscheidung zwischen der SWP und anderen Gesellschaften im Bereich der Politikberatungen, wie beispielsweise Lobbyisten, privatwirtschaftlichen Beratungsgesellschaften und Nichregierungsorganisationen, vgl. [33].

1.3. Deutsches Institut für Internationale Politik und Sicherheit

Das Deutsche Institut für Internationale Politik und Sicherheit der Stiftung Wissenschaft und Politik gliedert sich nach [47] in zwei zentrale Bereiche, den Forschungs- und den Fachinformationsbereich sowie einen zusätzlichen Servicebereich.

Der Forschungsbereich befasst sich mit der zentralen, politikwissenschaftlichen Arbeit des Instituts. Dieser Bereich gliedert sich in acht Forschungsgruppen, die sowohl thematische/prozeduale Komplexe darstellen (beispielsweise EU-Integration, EU-Außenbeziehungen), als auch Weltregionen abbilden (beispielsweise Asien und Amerika). Darüber hinaus gehören zu diesem Forschungsbereich auch mehrere Projektgruppen, die sich mit der Forschung in Drittmittel-geförderten Projekten befassen.

Der Fachinformationsbereich realisiert die informationswissenschaftlichen Aspekte der Dienstleistungen des Instituts. Dieser Bereich gliedert sich in das Informationsmanagement und ebenfalls acht Fachreferate, die ähnlich denen des Forschungsbereichs thematisch und geopolitisch strukturiert sind. Zuzüglich sind dem Fachinformationsbereich außerdem die Referate, die sich mit dem Fachinformationsverbund und der Bibliothek befassen, unterstellt.

Der Fachinformationsbereich der Stiftung Wissenschaft und Politik hat nach [46] drei zentrale Aufgabengebiete: Erstens die Realisierung der Informationsdienstleistungen für den Forschungsbereich, zweitens das Anbieten von Dienstleistungen als Fachinformationseinrichtung für den Bundestag und die Bundesministerien und drittens die Verwaltung und Entwicklung des Fachinformationsverbundes Internationale Beziehungen und Länderkunde (FIV).

1.4. Fachinformationsverbund

Die zugrundeliegenden Daten, für die das Information Retrieval System entwickelt und auf denen es evaluiert wird, wurden von der SWP zur Verfügung gestellt. Es handelt sich um einen umfassenden Auszug aus der Referenzdatenbank des Fachinformationsverbundes Internationale Beziehungen und Länderkunde. Im Folgenden soll kurz der Hintergrund des Fachinformationsverbundes geklärt, sowie ein Einblick in die Art der

Abbildung 1.1.: Kooperationspartner im Fachinformationsverbund

Datenbasis gegeben werden.

1.4.1. Entwicklung und Aufgaben des Fachinformationsverbundes

Die Grundlage für die spätere Entwicklung des Fachinformationsverbundes wurde 1978 durch die Kooperation zwischen der SWP und dem Bundesinstitut für ostwissenschaftliche und internationale Studien im Hinblick auf die gemeinsame Entwicklung der SWP-Datenbasis gelegt. 1984 wurden, im Rahmen eines Projektes, die Datenbanken der SWP und des Deutschen Übersee-Institut kombiniert. 1986 erfuhr der Fachinformationsverbund eine wichtige Bestätigung durch die Bundesregierung: Von diesem Zeitpunkt an wurde der FIV finanziell unterstützt, sodass sich weitere Institute dem Verbund anschließen konnten. 1992 wurde der erste öffentliche Zugang auf den Datenbestand des Verbundes unter dem Namen World Affairs Online eingerichtet. Ebenfalls in diesem Jahr begann die erste Zusammenarbeit auf europäischer Ebene im Rahmen des European Information Network on International Relations and Area Studies (EINIRAS). Im Statut von 1998 wurde die SWP als federführendes Institut im Fachinformationsverbund benannt. Im Jahr 2000 wurde die finanzielle Förderung von einer projektbasierten zu einer institutionellen Zuwendung gewandelt. Seitdem haben sich mehrere weitere Institute an der kontiuierlichen Entwicklung der Datenbasis beteiligt. Die Datenbasis wurde von EINIRAS-Kooperationspartnern auf den siebensprachigen europäischen Thesaurus erweitert. Eine über diese kurze Zusammenfassung hinausgehende Chronik findet sich in [42].

Der Fachinformationsverbund Internationale Beziehungen und Länderkunde stellt heute eine Kooperation zwischen zwölf deutschen Stiftungen, Instituten und Univer-

sitäten (vgl. Abbildung 1.1) dar, die in Fragen der Außen- und Sicherheitspolitik oder der Regionalwissenschaften forschen. Alle Kooperationspartner werden mehrheitlich durch die öffentliche Hand finanziert. Sie teilen in dieser Kooperation die Arbeit der Auswahl, Erschließung und Bereitstellung der Dokumente und haben gleichzeitig entsprechende Nutzungsrechte auf die geschaffene Datengrundlage. Darüber hinaus werden in Kooperation die Hilfsmittel wie Thesauri und Klassifikationen weiterentwickelt. Ziel des Fachinformationsverbundes ist zuvorderst die Informationsdienstleistung an Regierung, Ministerien und Bundestag. Nachgestelltes Ziel ist die Bereitstellung der Datenbasis für die Fachöffentlichkeit, vgl. [44].

1.4.2. Art der Datengrundlage des Fachinformationsverbundes

Bei der Datenbank des Fachinformationsverbundes handelt es sich um eine thematisch spezialisierte Form der Referenzdatenbank. Im Gegensatz zu anderen Datengrundlagen, wie beispielsweise Sammlungen von Volltexten, hat diese Form der Datenbasis zumeist die Eigenschaft, weitgehend durch Metainformationen erschlossen zu sein. Die zentrale Aufgabe der Datenbasis ist schließlich nicht, die tatsächlichen Volltexte vorzuhalten, sondern vielmehr eine bestimmte Menge an Artikeln, Büchern oder - ganz allgemein - Dokumenten zu verzeichnen und diese Dokumente möglichst treffend und eindeutig zu beschreiben. Damit verhält sich eine solche Referenzdatenbank wie ein gut organisiertes Bibliotheksverzeichnis, das dabei helfen kann, ein interessantes Buch schnell aufzufinden.

> „In bibliographic systems the use of subject headings or descriptors from a controlled vocabulary (a thesaurus or list of subject headings), to describe document content is common. By searching on a well designed and implemented controlled vocabulary, users can improve results.“ (vgl. [41])

Zusätzlich zu der klassischen bibliografischen Struktur, die eine solche Datengrundlage üblicherweise aufweist, verfügt die Datenbasis des FIV bei mehr als einem Viertel der verzeichneten Dokumente über Zusammenfassungen (vgl. Abschnitt 4.1.4 auf Seite 30). Die Kombination von kontrollierten Metainformationen und kurzen, die verzeichneten Dokumente beschreibenden Freitexten ist die Grundlage für das im Abschnitt 8.5 ab Seite 50 vorgestellte, auf das FIV angepasste Konzept eines Entry Vocabulary Moduls.

2. Information Retrieval

Das Informationszeitalter ist nur wenige Jahrzehnte alt und trotzdem hat bereits in diesen wenigen Jahren die Produktion und der Transport von Informationen Dimensionen angenommen, die zuvor kaum vorstellbar waren. Die Rate, in der Informationen in jeglicher Form generiert, übermittelt und gespeichert werden, war zu keinem anderen Zeitpunkt größer als sie heute ist. Entsprechend einer Untersuchung der University of California in Berkeley (vgl. [31]) hat sich die Menge der auf Papier, Film, magnetischen und optischen Speichermedien geschriebenen, neuen Daten innerhalb von drei Jahren (1999-2002) verdoppelt. Allein die Kommunikation per E-Mail bewirkte dieser Untersuchung zufolge einen Datenaustausch in der Größenordnung von 400 Millionen GigaByte im Jahr 2002. Diese Entwicklung ist offensichtlich direkt mit den positiven Entwicklungen von Bildungsstandards und Wohlstand in den Industrie- bzw. Postindustrienationen der Erde verknüpft. Während diese Faktoren die notwendige, gesellschaftliche Basis für das Informationszeitalter gelegt haben, hat der technische Fortschritt im Bereich der Informations- und Kommunikationstechnologien der Gesellschaft die dafür notwendigen Werkzeuge an die Hand gegeben. Die weite Verbreitung und gesellschaftliche Akzeptanz der Informationstechnologie hat denen, die an dieser Entwicklung teilnehmen, zuvor ungeahnte Möglichkeiten gegeben, sich zu informieren, zu kommunizieren, Informationen zu generieren und zu rezipieren.

Diese Fülle von Informationen hat allerdings nur dann einen Wert, wenn sie im annähernd gleichen Maße nutzbar gemacht wird. Dieser Aspekt der Nutzbarkeit kann sowohl durch bewußte Zugriffsbeschränkungen auf Informationseigentum (vgl. [29], Seite 12) als auch durch den Mangel an Zugang zu Informationen und Informationstechnologie auf einer entwicklungstechnischen Ebene (vgl. [29], Seite 136) sowie durch viele weitere Gründe beeinträchtigt werden.

Im Rahmen des Information Retrieval (IR) bezieht sich die Frage der Nutzbarkeit auf die Erschließung der Information innerhalb der Dokumente und Sammlungen von Daten.

> „Daten können aber nur genutzt werden, wenn sie auch erschlossen sind, wenn also diejenigen, die sie nutzen wollen oder sollen, auch wissen, dass und wo es die Daten gibt, wie sie gesuchte Informationen darin finden können und wie sie diese nutzen können und dürfen.“ (vgl. [15], Seite 3)

Diese Erschließung von Informationsbeständen nach ihren Inhalten, das Information Retrieval, ist dementsprechend eine der zentralen Aufgaben in einer Informationsgesellschaft, die immer mehr Informationen produziert, auf Informationen angewiesen ist und von der Nutzbarkeit der Informationen abhängt.

Die Fachgruppe Information Retrieval der Deutschen Gesellschaft für Informatik definiert IR wie folgt:

> „Im Information Retrieval (IR) werden Informationssysteme in bezug auf ihre Rolle im Prozeß des Wissenstransfers vom menschlichen Wissensproduzenten zum Informations-Nachfragenden betrachtet. Die Fachgruppe „Information Retrieval“ in der Gesellschaft für Informatik beschäftigt sich dabei schwerpunktmäßig mit jenen Fragestellungen, die im Zusammenhang mit vagen Anfragen und unsicherem Wissen entstehen. Vage Anfragen sind dadurch gekennzeichnet, daß die Antwort a priori nicht eindeutig definiert ist. [...] Die Unsicherheit (oder die Unvollständigkeit) dieses [des gespeicherten, Anm. d. Aut.] Wissens resultiert meist aus der begrenzten Repräsentation von dessen Semantik (z.B. bei Texten oder multimedialen Dokumenten); darüber hinaus werden auch solche Anwendungen betrachtet, bei denen die gespeicherten Daten selbst unsicher oder unvollständig sind (wie z.B. bei vielen technisch-wissenschaftlichen Datensammlungen). Aus dieser Problematik ergibt sich die Notwendigkeit zur Bewertung der Qualität der Antworten eines Informationssystems, wobei in einem weiteren Sinne die Effektivität des Systems in bezug auf die Unterstützung des Benutzers bei der Lösung seines Anwendungsproblems beurteilt werden sollte.“ (vgl. [16])

Aus dieser Definition für Information Retrieval lassen sich vier zentrale Charakteristiken des IR zusammenfassen:

- Information Retrieval befasst sich mit dem Prozess des Wissenstransfers.
- Die Anfragen im Rahmen des Wissenstransfers sind zumeist vage und ungenau definiert.
- Die Wissensgrundlage besteht aus ungenau definierten Wissenseinheiten.

- Die durch die Ungenauigkeit in Anfrage und Wissensbestand gegebene Unsicherheit erfordert eine besondere Form der Evaluierung von Systemen (Orientierung am Nutzen der Ergebnisse für den Anwender).

Die zentrale Gegenbenheit im Umgang mit Information Retrieval, die Unschärfe von Anfrage und Wissen, grenzen dieses Feld beispielsweise vom Data Retrieval (DR) ab. DR ist vielmehr das Zusammenstellen einer Selektion von Dokumenten, die exakt einem vorgegebenen Muster entsprechen, während die Aufgabe des IR das Auffinden von Dokumenten mit relevanten Informationen für ein vorgegebenes Thema bzw. einer Anfrage ist.

Während sich Data Retrieval also durch seine Klarheit und Beschränkung auf das Zusammenstellen von Selektionen aus Objekten auszeichnet, beschreibt das Information Retrieval, wie weiter oben dagelegt, eine Methode des informationellen Wissenstransfers: Data Retrieval verwendet zum Abgleichen einer Anfrage mit der Datengrundlage das Konzept exakter Übereinstimmung, während IR versucht, die Relevanz eines Dokuments im Kontext einer Anfrage zu bewerten. IR versucht in diesem Bezug nicht nur eine durch die Anfrage vorgegebene Signatur, sondern vielmehr die im semantischen Sinne relevanten Dokumente aufzufinden. Einen Überblick über viele eindeutige Unterscheidungen zwischen DR und IR hat Keith Rijsbergen in [39] vorgestellt, siehe Tabelle 2.1.

	Data Retrieval	Information Retrieval
Matching	Exact match	Partial match, best match
Inference	Deduction	Induction
Model	Deterministic	Probablistic
Classification	Monothetic	Polythetic
Query language	Artificial	Natural
Query specification	Complete	Incomplete
Items wanted	Matching	Relevant
Error response	Sensitive	Insensitive

Tabelle 2.1.: Data Retrieval vs. Information Retrieval nach Keith Rijsbergen in [39]

2.1. Beispiele für Retrieval-Modelle

Es gibt mehrere Modelle, in denen Retrieval Systeme umgesetzt werden können. Die im Folgenden beschriebenen, klassischen Modelle stellen nur eine grundlegende Auswahl dar. Viele Systeme basieren auf einer Mischform von mehreren Retrievalmodellen, um die Retrievalqualität der Systeme zu steigern. Dabei wird beispielsweise die boolesche

Logik mit einer probablistischen Methode zur Anfragenerweiterung durch potentiell nützliche Terme und zur Bewertung der Relevanz der Ergebnisse kombiniert.

2.1.1. Boolesches Retrieval-Modell

Das Boolesche Retrieval ist streng genommen eine Form des Data Retrieval. Es beschränkt sich darauf, eine Anfrage mit der verzeichneten Datengrundlage exakt abzugleichen und im Falle der Übereinstimmung die gefundenen Dokumente unsortiert zurückzugeben. Es lässt sich also als eine einfache, binäre Bewertung von Dokumenten nach einem Kriterium oder mehreren Kriterien beschreiben. Dabei basiert das Grundkonzept des Modells auf der objektweisen Verknüpfung von Attribut und Wert. Alle in einer Datengrundlage verzeichneten Objekte werden auf diese Art und Weise durch Attribute schematisiert und durch die Ausprägungen dieser Attribute unterschieden. Eine Anfrage wird entsprechend in Paaren von Attribut und Wert formuliert. Im Rahmen des Booleschen Retrievalverfahrens kann darüber hinaus eine logische Verknüpfung von Anfragetermen angewendet werden. Die Operatoren AND, OR und NOT bieten hierbei die Möglichkeit, Objekte, bei denen sowohl die eine als auch die andere Ausprägung eines Attributs gefunden werden konnte (AND), bei denen entweder die eine oder die andere Ausprägung eines Attributs (OR) oder aber die eine aber nicht die andere Ausprägung eines Attributs (NOT), auszuwählen.

Nach Baeza-Yates und Ribeiro-Neto ist das Boolesche Modell formal wie folgt definiert:

> „For the Boolean model, the index term weight variables are all binary, i.e., $\omega_{i,j} \in \{0, 1\}$. A query q is a conventional Boolean expression. Let $\vec{q}_{dnf}$ be the disjunctive normal form for the query q. Further, let $\vec{q}_{cc}$ be any of the conjunctive components of $\vec{q}_{dnf}$. The similarity of a document d_j to the query q is defined as
>
> $$\mathbf{sim(d_j, q)} = \begin{cases} 1 & if \quad \exists \vec{q}_{cc} | (\vec{q}_{cc} \in \vec{q}_{dnf}) \cap (\forall_{k_i}, g_i(\vec{d_j}) = g_i(\vec{q}_{cc})) \\ 0 & otherwise \end{cases}$$
>
> If $sim(d_j, q) = 1$, then the Boolean model predicts that the document d_j is relevant to the query q (it might not be). Otherwise, the predicition is that the document is not relevant.“ (vgl. [4], Seite 26)

In der Booleschen Syntax lassen sich dabei auch kompliziertere Anfragekonstrukte miteinander kombinieren und verschachteln (vgl. Beispiel auf Seite 33).

Die Nachteile des Verfahrens liegen darin, dass diese Form des Daten Retrieval die Ergebnisse nicht nach potentieller Relevanz sortiert, da es nur binär zwischen Übereinstimmung und Differenz unterscheiden kann. Zwei weitere Nachteile sind, dass die Syntax der Anfrage einerseits nicht natürlichsprachlich ist und andererseits ggf. ein Wissen über die Systematik des Aufbaus der Datengrundlage (Bezeichnung von Attributen) voraussetzt. Außerdem ergibt sich im Rahmen des Booleschen Retrieval oftmals das Problem, dass die Gruppe der gefundenen Dokumente meist deutlich zu groß (potentiell großer Anteil an irrelevanten Dokumenten) oder zu klein (exakte Anfrage führte zur Ausgrenzung von ebenfalls relevanten Dokumenten, die mit anderen Attributen beschrieben wurden) ist.

2.1.2. Vektorraum-Modell

Das Vektorraum-Modell bewertet die Ähnlichkeit bestimmter Dokumente zu einer Anfrage. Jedem Dokument in einem Datenbestand wird aufgrund der Ausprägung seiner Attribute ein Vektor im n-dimensionalen Raum zugeordnet, wobei n gleich der Anzahl der Attribute multipliziert mit allen Ausprägungen ist. Bei einem reinen Volltextdokument entspricht n der Anzahl der verschiedenen Terme im Dokument.

Zentrale Maßzahlen für die Berechnung der Gewichte von Termen sind zum einen die Häufigkeit des Auftretens eines Terms in einem Dokument, der Termfrequenz, und zum anderen die Anzahl der Dokumente, in denen der Term auftritt: die invertierte Dokumentfrequenz. In der klassischen Form ergibt das Produkt aus normalisierter Termfrequenz

$$ntf = \frac{\text{Anzahl Nennungen Term } t \text{ in Dokument } d}{\text{Anzahl aller Terme in } d}$$

und invertierter Dokumentfrequenz

$$idf = log \frac{\text{Anzahl aller Dokumente}}{\text{Anzahl aller Dokumente mit } t}$$

das Gewicht des Terms.

Anfrageseitig wird in Relation zu den in der Anfrage beinhalteten Terme der Vektor der Anfrage berechnet und im Rankingprozess der Vektor der Anfrage mit den Vektoren aller Objekte verglichen und entsprechend der Ähnlichkeit sortiert. Ziel ist,

durch die Ähnlichkeit eines Anfragevektors und eines Objektvektors die Relevanz beider Elemente füreinander zu bewerten.

Nach Ferber ist das Vektorraum-Modell mit Attributen formal wie folgt definiert:

> „Sei $D = \{d_1, ..., d_m\}$ eine Menge von Dokumenten oder Objekten und $A = \{A_1, ..., A_n\}$ eine Menge von Attributen $A_j : D \to \mathbf{R}$ auf diesen Objekten. Die Attributwerte $A_j(d_i) =: \omega_{i,j}$ des Dokuments d_i lassen sich als Gewichte auffassen und zu einem Vektor $\omega_i = (\omega_{i,1}, ..., \omega_{i,n}) \in \mathbf{R}^n$ zusammenfassen. Dieser Vektor beschreibt das Dokument im Vektorraummodell: Er ist seine Repräsentation und wird Dokumentvektor genannt. Eine Anfrage wird durch einen Vektor $q \in \mathbf{R}^n$ mit Attributwerten, den Anfragevektor oder Query-Vektor, dargestellt. Eine Ähnlichkeitsfunktion $s : \mathbf{R}^n \times \mathbf{R}^n \to \mathbf{R}$ definiere für je zwei Vektoren $x, y, \in \mathbf{R}^n$ einen reellen Ähnlichkeitswert $s(x, y)$.“ (vgl. [15], Seite 62)

Um die Ähnlichkeit zwischen den Dokumentvektoren und dem Anfragevektor zu bewerten, kann die folgende Funktion angewendet werden, die die Differenz zwischen den Vektoren der Anfrage und des Dokuments berechnet.

$$sim(d_j, q) = \frac{\vec{\omega_i} \bullet \vec{q}}{|\vec{\omega_i}| \times |\vec{q}|} = \frac{\sum_{i=1}^{t} \omega_{i,j} \times \omega_{i,q}}{\sqrt{\sum_{i=1}^{t} \omega_{i,j}^2} \times \sqrt{\sum_{j=1}^{t} \omega_{i,q}^2}}$$

Das Vektormodell hat gegenüber dem Booleschen Modell den großen Vorteil, dass das binäre durch ein graduelles Bewertungssystem ersetzt wird. Es gibt eine weite Bandbreite zwischen der maximalen Bewertung von relevant (Bewertung mit 1) und irrelevant (Bewertung mit 0) die, berechnet für jedes einzelne gefundene Dokument im Kontext der Anfrage, zur Erstellung einer Reihenfolge der Dokumente nach Ähnlichkeit zur Anfrage verwendet werden kann.

2.1.3. Probablistisches Modell

Das probablistische Retrievalmodell basiert auf Berechnungen aus der Wahrscheinlichkeitsrechnung. Das zentrale Ziel der Bewertung eines Dokuments im Kontext einer gegebenen Anfrage ist die Berechnung der Relevanzwahrscheinlichkeit. Hierbei dient eine Analyse der Verteilung der Termen in der Dokumentsammlung und den einzelnen Dokumenten als ein Indikator für die Relevanz des Dokuments. Je höher die Wahrscheinlichkeit der Relevanz, desto höher wird ein Dokument auf einer entsprechend sortierten Ergebnisliste aufgeführt.

Zur Berechnung der Relevanzwahrscheinlichkeit kann beispielsweise der Retrieval Status Value (RSV) verwendet werden, der wie folgt formal definiert ist:

$$rsv = \sum_{\{i \in I | t_i \in q \cap d\}} \left(log \frac{r_i}{n_i} + log \frac{(1 - n_i)}{(1 - r_i)} \right)$$

> „Um das Verfahren anzuwenden, müssen Werte für r_i und n_i geschätzt werden. r_i ist die Wahrscheinlichkeit, dass der Term t_i in einem für die Anfrage q relevanten Dokument vorkommt, und n_i ist die Wahrscheinlichkeit, dass der Term t_i in einem für die Anfrage q nicht relevanten Dokument vorkommt.“ (vgl. [15], Seite 191)

Zentrale Annahme für die Berechnung des Retrieval Status Value ist hierbei, dass das Auftreten verschiedener Terme in Dokumenten unabhängig voneinander ist. Ein Problem bei der Berechnung des RSV ist, dass die Anzahl von relevanten Dokumenten im Vorfeld der Berechnung geschätzt werden muss.

Weitere Maßzahlen zur Berechnung der Relevanzwahrscheinlichkeit sind die Bewertung der Termgewichte nach der Robertson-Sparck-Jones Formel oder auch die Berechnung durch die aus dieser Formel entwickelten Okapi-Formel. Die Berechnung der Relevanzwahrscheinlichkeit erfolgt dann beispielsweise über die Berechnung des Skalarprodukts des Vektors der Anfrage mit den Vektoren der einzelnen Dokumente.

Ein weitergehender Einblick in die probablistischen Retrievalansätze findet sich in [39], Seiten 87-95, sowie in [10].

2.2. Hilfsmittel zur Verbesserung der Retrievalqualität

Im Rahmen des Information Retrieval gibt es viele verschiedene Ansätze zur Verbesserung von Retrievalergebnissen. Darunter sind Methoden zur Veränderung der Morphologie der Anfrageterme (Grundformenreduktion, Kompositazerlegung, vgl. [8] und [1]), die Erkennung und Verknüpfung von Phrasen in Anfragen (vgl. [18] und [30]), die Entfernung von nicht aussagekräftigen Termen in der Anfrage (Stoppwortlisten) und viele weitere. Im Folgenden sollen zwei Hilfsmittel vorgestellt werden, deren Konzepte als Grundlage für den Ansatz des in Abschnitt 5 vorgestellten Systems dienen.

2.2.1. Klassifikationen und Thesauri

Das Konzept der Klassifikation ist in jeder Bibliothek systematisch abgebildet und hat sich somit zumindest in der Domäne des gedruckten Wissensbestands durchgesetzt. Die Klassifikation eines Datenbestands bedeutet, ihn in einzelne Klassen zu fragmentieren. In einer Bibliothek geschieht das zunächst durch die thematische Aufteilung der Dokumente in Fachbereiche und darauf folgend in Themengebiete, Spezialgebiete und so weiter. Es lässt sich einfach erkennen, dass durch eine solche mehrstufige Klassifikation ein hierarchisches System entsteht, welches sich von der obersten zur untersten Ebene thematisch immer weiter spezialisiert und detailliert. Auf der untersten Ebene finden sich dementsprechend im Idealfall Kategorien, die sich nicht weiter aufteilen lassen. Um eine Klassifikation zu entwickeln, ist entsprechend eine Definition und Systematisierung der Klassen notwendig. Zentraler Aspekt bei dieser Systematisierung der Klassen ist, dass die Klassifikation trennscharf entwickelt wird und die zugrundeliegenden Objekte eindeutig (mindestens) einer Kategorie zugeordnet werden können. Je nachdem wie konsequent die Hierarchie umgesetzt wird, kann man neben starken Hierarchien (jedes Dokument eindeutig in einer Kategorie) auch schwache Hierarchien erzeugen (einzelne Dokumente in mehreren Kategorien, einige Kategorien mit mehreren übergeordneten Kategorien auf der nächsthöheren Ebene). Die Zugehörigkeit eines Objekts in einer bestimmten Kategorie lässt sich in Form eines Attributs dieses Objekts ausdrücken und auf diese Art und Weise wiederum im Information Retrieval Prozess verwenden.

Beispiele für hierarchische Klassifikationssysteme sind sowohl die Internationale Dezimalklassifikation [35] als auch das in Kapitel 4.1.4 auf Seite 30 beschriebene Klassifikationsystem des Fachinformationsverbundes.

Während Klassifikationen die Dokumente einer Grundgesamtheit in Kategorien

hierarchisch einteilen, kann ein Thesaurus einen weiteren wichtigen Beitrag zur inhaltlichen Erschließung der Inhalte von Dokumenten leisten: In Attributen an Objekte vergebene und in Thesauri verzeichnete Deskriptoren beschreiben den Inhalt des Dokuments eindeutig durch ihre kontrollierte Formulierung und Vergabe. Das kontrollierte Vokabular hilft durch die geringere Anzahl und sorgfältige Auswahl von geeigneten Vokabeln dabei, die Entscheidung, ob ein Objekt relevant oder irrelevant ist, eindeutiger zu machen und unter anderem das klassische Problem der Synonymie (zumindest innerhalb der Sammlung von Deskriptoren) zu lösen. Die zentrale Aufgabe eines Thesaurus ist neben der Erstellung einer kontrollierten Deskriptorenterminologie die Verknüpfung von Begriffen mit weiteren, semantisch verwandten Begriffen (vgl. die Untersuchungen durch Voorhees in [49]). Ergänzend zu Synonymen können außerdem Antonyme, Oberbegriffe oder detailliertere Begriffe sowie auch Übersetzungen aufgeführt werden. Auf diese Weise lassen sich beispielsweise Anfragen an ein IR-System durch verwandte Begriffe erweitern und im Suchprozess Dokumente, die durch Antonyme beschrieben werden, ausschließen. Durch Ober- und Unterbegriffe in Thesauri kann sich wie in der Klassifikation eine hierarchische Struktur ergeben.

Sowohl Klassifikationsangaben als auch Deskriptoren können als Metadaten eines entsprechend erschlossenen Dokuments bezeichnet werden. Zu solchen Metadaten zählen besonders in Referenzdatenbanken noch eine Vielzahl weiterer Informationen wie bspw. die Angabe von Autoren, das Erscheinungsdatum und so weiter. Allerdings werden sowohl Klassifikationsangaben als auch Deskriptoren üblicherweise für die gegebene Datenbank entwickelt, für jedes Dokument (entweder automatisch oder manuell) erarbeitet und vergeben und sind somit Teil der Entwicklungsarbeit eines Datenbestands. In dieser Arbeit und dem vorgestellten System werden sowohl Klassifikationsangaben als auch die beschreibenden Terme und Phrasen des Thesaurus effektiv als Deskriptoren für Dokumente verwendet, da das vorgestellte System im aktuellen Entwicklungsstadium nicht die Hierarchie der Klassifikation berücksichtigt.

2.2.2. Relevance Feedback

Relevance Feedback ist eine Methode zur Anpassung der Suchanfrage im Sinne der höheren Effektivität und der besseren Retrievalqualität insgesamt. Dieses Verfahren lässt sich sowohl in Interaktion mit dem Nutzer des Information Retrieval Systems realisieren als auch durch ein automatisiertes Verfahren (vgl. beispielsweise [32]).

Bei der interaktiven Variante des Relevance Feedback wählt der Nutzer des Systems aus der Auflistung der mit Hilfe der ursprünglichen Anfrage gefundenen Dokumente diejenigen Dokumente aus, die am ehesten relevant für seine Suchanfrage sind. Auf Basis der Terme in diesem Dokument wird daraufhin die Suchanfrage so verändert,

dass sie die als relevant erkannten Dokumente besser beschreibt und diese damit höher bewertet, sowie weitere, relevante Dokumente findet.

Im Rahmen der automatischen Variante des Relevance Feedback analysiert das System selbständig eine gewisse Anzahl der am besten bewerteten (und damit als relevant angenommenen) Dokumente aus einer vorläufigen Suche, ergänzt die ursprüngliche Suchanfrage um die geeignetsten Terme und führt den Suchvorgang erneut aus.

Formal wird diese Methode nach Rocchio in [23] wie folgt beschrieben:

> „After the user responds, the set R contains the n_1 relevant document vectors, and the set S contains the n_2 non-relevant document vectors. Rocchio builds the new query Q' from the old query Q using the equation given below:
>
> $$Q' = q + \frac{1}{n_1}\sum_{i=1}^{n_1} R_i - \frac{1}{n_2}\sum_{i=1}^{n_2} S_i$$
>
> R_i and S_i are individual components of R and S respectively.“ (vgl. [23], Seite 96)

Es existieren mehrere abgewandelte Modelle, bei denen den einzelnen Vektoren verschiedene Gewichte zugewiesen werden können. Mit diesem Ansatz wird versucht zu erreichen, dass durch die Ergänzung der Anfrage um relevante Terme die Gewichtung der scheinbar ungeeigneten Terme abnimmt und so die Retrievalqualität insgesamt zunimmt. Attar und Fraenkel hatten bereits in [3] das Potential von Blind Relevance Feedback detailliert vorgestellt. Relevance Feedback kann sowohl, wie hier vorgestellt, im Vektorraum-Modell als auch im probablistischen Retrieval-Modell umgesetzt werden, einen aktuellen Überblick bietet [23] auf Seite 94 bis 105.

2.3. Aspekte des mehrsprachigen Information Retrieval

Die Internationalisierung von Datenbeständen und deren Nutzerschaft hat es im Laufe der vergangenen Jahre notwendig gemacht, Informationen auch unabhängig von Sprachbarrieren auffindbar und damit nutzbar zu machen. Mit Hilfe der mehrsprachigen Funktionalität eines Information Retrieval Systems kann der Nutzer in einer ihm geläufigen Sprache eine Anfrage formulieren und eine Ergebnisauflistung aus Dokumenten aller möglichen Sprachen erhalten.

> „Successful cross-language information retrieval combines linguistic techniques (phrase discovery, machine translation, bilingual dictionary lookup) with robust monolingual information retrieval.“ (vgl. [20])

Besonders im Hinblick auf die Funktionen eines Information Retrieval Systems für Referenzdatenbanken mit Dokumenten von internationalem, thematischen Bezug kann eine solche Funktion offensichtlich sehr nützlich sein. Natürlich ist bei dem Einsatz und der Benutzung einer mehrsprachigen Funktionalität eines Information Retrieval Systems immer vorausgesetzt, dass der Nutzer die Sprachen, in denen die Dokumente verfasst sind, zumindest ansatzweise beherrscht, da sonst ein Grundziel der Leistung des Information Retrieval Systems, die Assistenz des Nutzers, nicht erfüllt werden könnte. Streng genommen müsste das System in einem solchen Fall jedes vermeintlich relevante Dokument im Rahmen des Retrievalprozesses zusätzlich übersetzen, um seinen Zweck zu erfüllen.

Um eine mehrsprachige Retrievalleistung zu erzielen, gibt es zwei klassische Ansatzpunkte: Einerseits die Anfragenübersetzung in die jeweilige Dokumentsprache und andererseits die Übersetzung der im Index verzeichneten Dokumente in die potentielle Anfragesprache. Beide Ansätze können sowohl auf der Basis der Termübersetzung als auch der Übersetzung ganzer Phrasen umgesetzt werden.

Es gibt mehrere Ansätze zur Realisierung von mehrsprachigen Retrieval Systemen, darunter sind Konzepte, die Parallelkorpora (Sammlungen der selben Dokumente in mehreren Sprachen) oder vergleichbare Korpora (Sammlungen von thematisch ähnlichen Dokumenten in verschiedenen Sprachen) verwenden. Darüber hinaus gibt es mehrere Methoden zur Verbesserung der Ergebnisse, wie auf bestimmte Sprachen spezialisierte Grundformenreduktion und die Erkennung und Transliteration von Eigennamen.

Einen umfassenden Überblick über die vielen verschiedenen Methoden und Strategien im Rahmen des mehrsprachigen Information Retrieval bietet [23], Seiten 149-179.

2.4. Evaluierung von Information Retrieval Systemen

Die Evaluierung eines Systems bedeutet, es durch eine objektive Bewertung an einer Skala entweder im umfassenden Sinne der Leistungsfähigkeit oder im Sinn eines Teilbereichs der Leistungsfähigkeit (Kosten, Geschwindigkeit, Ergebnisqualität, etc.) zu messen. Eine Evaluierung setzt klare Kriterien und eindeutige Maßzahlen voraus. In vielen Teilbereichen der Systementwicklung ist eine solche Quantifizierung möglich

(Preis, Geschwindigkeit). In der zentralen Disziplin der Retrievalqualität ist aufgrund der Ungewissheit der Relevanz eines Dokuments für eine Anfrage eine eindeutigen Aussage über die Qualität des Ergebnisses schwieriger.

Entsprechend der Definition der Fachgruppe für Information Retrieval (vgl. Seite 8 und [16]) soll „die Unterstützung des Benutzers bei der Lösung seines Anwendungsproblems“ beurteilt werden. Da sich eine solche Bewertung nicht für jede mögliche Anfrage in Kombination mit jedem verzeichneten Dokument prüfen lässt, werden zum Zweck der Untersuchung der Leistungsfähigkeit von IR Systemen üblicherweise Ausschnitte von Datengrundlagen nach Relevanz für eine Anzahl von - für die Anwendungssituation exemplarischen- Anfragen bewertet.

Das Konzept der Relevanz ist von zentraler Bedeutung für die Evaluierungsmethoden und -maßstäbe für IR Systeme. Die Bewertung der Relevanz von gefundenen Dokumenten gegenüber der eingegebenen Anfrage ist ebenfalls das zentrale Problem der Evaluierungssystematik: Es gibt keine einheitliche Definition für die Relevanz eines Dokuments für eine Anfrage. Im Sinne des definitorischen Bezugs der Leistung des Systems auf die Unterstützung des Nutzers werden daher üblicherweise fachlich qualifizierte Personen (d.h. Personen, die sich in die Situation des Nutzers mit einer gegebenen Anfrage versetzen und dabei den Informationsbedarf abschätzen können) damit beauftragt, die Relevanz eines Dokuments für eine Anfrage auf Basis Ihrer Fachkenntnisse zu bewerten.

Nach der Beurteilung der in einem Evaluierungsprojekt mit gegebene Anfragen gefundenen Dokumente lassen sich in Zusammenhang mit den Ergebnisauflistungen der IR-Systeme Maßzahlen errechnen, die die Leistungsfähigkeit des Systems abbilden können. Die am häufigsten verwendeten Maßzahlen zur Bewertung der Retrievalqualität von IR Systemen sind Recall und Precision.

$$recall = \frac{\text{Anzahl der gefundenen, relevanten Dokumente}}{\text{Anzahl aller relevanten Dokumente}}$$

$$precision = \frac{\text{Anzahl der gefundenen, relevanten Dokumente}}{\text{Anzahl aller gefundenen Dokumente}}$$

Entsprechend fungiert Recall als der Wert, der die Retrievalqualität mit Blick auf das Auffinden aller relevanten Dokumente bewertet, während Precision den Anteil des ebenfalls gefundenen Ballasts an irrelevanten Dokumenten in seine Bewertung mit

einbezieht. Für beide Maßzahlen ergibt sich ein Spektrum zwischen 0 (schlechtestes Ergebnis) und 1 (bestes Ergebnis).

Recall und Precision werden oft in Form eines Diagramms abgebildet. Die grundlegende Form des Recall-Precision Diagramms ist eine Punktwolke (vgl. Abbildung 2.1), bei der nach jedem aufgelisteten Dokument erneut Recall und Precision bewertet werden. Üblicherweise werden jedoch über elf Recall-Stufen (0,0, 0,1, 0,2 ... 1,0), die jeweils einen entsprechenden prozentualen Anteil der gefundenen, relevanten Dokumente repräsentieren, der durchschnittliche Werte der einzelnen, nach jedem Dokument berechneten Precision-Werte abgebildet und diese Precision-Mittelwerte durch eine Linie verbunden (vgl. Abbildungen wie beispielsweise 10.1 auf Seite 73). Formal entsteht dadurch kein Graph, da die Punkte auf den Linien zwischen den Mittelwerten nicht definiert sind. Ein Überblick über die gängigen Evaluierungsmaße findet sich in [48].

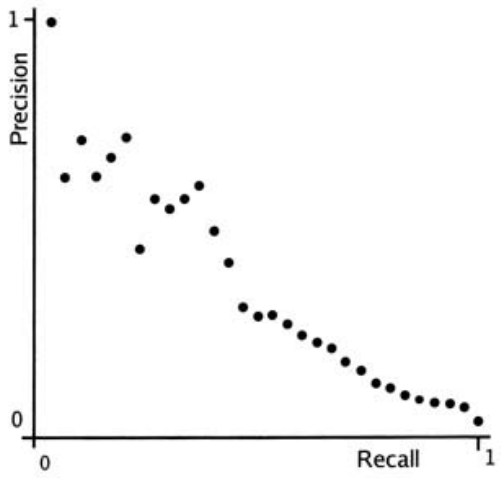

Abbildung 2.1.: Recall-Precision Diagramm aus [15], Seite 89

Es gibt mehrere Foren, in denen neue Ansätze zum Information Retrieval entwickelt und getestet werden. Hierzu gehört als spezialisiertes Forum für mehrsprachiges Retrieval das Cross Language Evaluation Forum CLEF[1] und als allgemeinere Plattform die Text Retrieval Conference TREC[2]. In beiden Foren arbeiten die beteiligten Forscher in verschiedenen spezialisierten Forschungsfeldern, die unterschiedliche Aspekte des Information Retrieval weiterentwickeln.

[1] vgl. auch http://www.clef-campaign.org
[2] vgl. auch http://trec.nist.gov

3. Entry Vocabulary

3.1. Die zentrale Frage des Vokabulars

Bei der Verbalisierung von Informationen wird die Nachricht mit Hilfe eines Vokabulars kodiert. Da es aufgrund verschiedener Sprachen und spezialisierter Fachsprachen viele verschiedene Vokabulare gibt, ist es essentiell, dass, sofern die Information ausgetauscht werden soll, sowohl der Sender als auch der Empfänger der Information dasselbe Vokabular beherrschen und den Sinn der verbalen Abbildung der Information verstehen können. Ist das in der Kommunikation verwendete Vokabular einem der Kommunikationspartner nicht bekannt, wird der Austausch von Informationen nahezu unmöglich.

Bezogen auf das Information Retrieval ergibt sich in diesem Kontext ein ähnliches Problem. Je nach Aufgabe und Einsatzgebiet des IR-Systems variiert die Art der Datengrundlage und der in der Datengrundlage verzeichneten Informationen drastisch. Handelt es sich um eine hochspezialisierte Datenbank, beispielsweise die in [17] herangezogene Datenbank von amerikanischen Im- und Exportstatistiken, so wird auch die Information in der Datengrundlage entsprechend in einem spezialisierten Vokabular kodiert sein. Im Falle der Außenhandelsstatistiken lässt sich beispielsweise nicht erfolgreich mit dem Begriff „automobile“ suchen - der entsprechende Begriff lautet in dem Zielvokabular des IR-Systems „Pass Mtr Veh“, was einen Abkürzung für „Passenger Motor Vehicle“ darstellt. Ein System, das den Vokabularunterschied zwischen dem Ausgangsvokabular des Nutzers und dem Eingangsvokabular der Datenbasis ausgleicht, ist in diesem Moment notwendig und wird üblicherweise realisiert durch, wo vorhanden und zugänglich, zu konsultierende Auflistungen des Datenbankvokabulars. Ist das Vokabular unbekannt und eine solche Auflistung des Datenvokabulars nicht zugänglich, wird die Datenbasis nicht (effektiv) zu durchsuchen sein.

Von zentralem Interesse für diese Arbeit ist der Faktor der Kontrolle des Vokabulars.

> „vocabulary control: The deliberate restriction of the type and number of words for [...] lexicographic purposes.
>
> vocabulary selection: ⇒ vocabulary control“ (vgl. [25])

Das Vokabular einer Sprache besteht aus allen Worten, die in dieser Sprache verbalisiert oder verschriftlicht werden können. Gemeinsprachliche Texte, die in dieser Sprache verfasst sind, verwenden dabei den gemeinsprachlichen Anteil, während fachlich spezialisierte Texte die entsprechende Fachterminologie ergänzen und somit den Vorteil der exakteren Beschreibungsmöglichkeiten durch Fachbegriffe nutzen. Fachtexte bestehen dabei immer aus einer Mischung von gemeinsprachlichem und fachsprachlichem Anteil an Vokabular (vgl. [2], Seite 188). Diese Wortschätze werden grundsätzlich als freie, d.h. nicht kontrollierte, Vokabulare bezeichnet.

Eine Terminologie von Deskriptoren ist dagegen eine vollständig kontrollierte Form eines Vokabulars, da eine Sammlung von Deskriptoren zum Zweck der eindeutigen Beschreibung von Objekten systematisch erarbeitet wird und damit in ihrem Umfang begrenzt sein muss. Zu einem solchen kontrollierten Vokabular von Deskriptoren gehört ebenfalls der Begriff „Pass Mtr Veh“ aus Geys Beispiel. An diesem Beispiel lässt sich außerdem zeigen, dass eine Form des kontrollierten Vokabulars nicht zwangsläufig aus Teilen des freien Vokabulars bestehen muss: Je nach Systematik können auch Abkürzungen oder Zeichencodes als Deskriptoren verwendet werden, die mehr oder weniger kryptisch wirken können.

Zwischen diesen beiden Extremen des gänzlich freien Vokabulars einer Sprache und des stark kontrollierten Vokabulars einer Deskriptorenterminologie befindet sich das Vokabular, in dem Volltexte oder Zusammenfassungen in Referenzdatenbanken geschrieben wurden. Das Vokabular wird durch die spezifische Wortwahl der Verfasser, die entsprechende Fachsprache, begrenzt und darüber hinaus durch Techniken wie Stoppwortlisten kontrolliert, die besonders häufig auftretende Terme entfernen. Dieses Vokabular ist offensichtlich ebenfalls zu einem gewissen Maße kontrolliert, hat aber deutlich freiere Züge als eine Sammlung von Deskriptoren. Daher wird im weiteren Verlauf dieser Arbeit bei indexierten Zusammenfassungen oder Titlen von im Vergleich zu den Deskriptoren freien oder „relativ freien“ Texten gesprochen.

Üblicherweise würde die von Gey beschriebene, besonders spezialisierte Datengrundlage nur für die Spezialisten zugänglich sein, die mit dem entsprechenden Vokabular der Datengrundlage vertraut sind. Allerdings kann es sein, dass auch eine solche Datenbank öffentlich zugänglich gemacht und somit auch von Nutzern durchsucht wird, die das spezifische Vokabular nicht kennen. Hierbei entsteht die Situation, dass Nutzer, die des Vokabulars des Systems nicht mächtig sind, das System nicht auf eine zielführende Art und Weise bedienen können - nicht nur, weil sie das Ergebnis des Retrievalprozess eventuell nicht interpretieren, sondern weil auf das System unvorbereitete Nutzer ohnehin kaum eine sinnvolle Anfrage formulieren können.

> „Searching of databases, textual or numeric, is likely to be effective and efficient only if the user is familiar with the classification, categorizing and indexing schemes (metadata vocabularies) being searched. Therefore, it is obviously beneficial to provide a bridge between the user's ordinary language and the metadata vocabularies of the unfamiliar database in order to compensate for abbreviated, cryptic or specialized terminologies." (vgl. [19])

Für die Lösung dieses Problems der semantischen Heterogenität in verschiedenen Metadatensystemen existieren mehrere Ansätze (vgl. [27]). Um aber eine Brücke zwischen dem spezifischen, kontrollierten Vokabular einer spezialisierten Datengrundlage und dem mehr oder weniger freien Vokabular eines unerfahrenen Nutzers zu bauen, wurden in den letzten Jahren in mehreren Projekten sogenannte Entry Vocabulary Module eingesetzt (vgl. [9]). Diese Module bestehen üblicherweise aus einem Entry Vocabulary Index, der die Beziehungen zwischen Termen des Freitexts und Deskriptoren oder Klassifikationsangaben auf Basis von Wahrscheinlichkeiten abbildet und einer Schnittstelle, die geeignete kontrollierte Vokabeln vorschlagen kann (vgl. [34]). Auf diese Weise kann für eine Anfrage, die frei formuliert wurde, passendes, kontrolliertes Vokabular zur Ergänzung oder Überarbeitung der Anfrage empfohlen werden.

Eine weitere, interessante Anwendungsmöglichkeit besteht außerdem darin, nicht nur einen „vertikalen" Vokabularunterschied zu nivellieren, sondern auch einen „horizontalen": Während der Unterschied zwischen spezialisiertem und freiem Vokabular eindeutig ist, ist der Unterschied zwischen dem Vokabular verschiedener Sprachen - also der mehrsprachige Aspekt - durch den Einsatz von EVMs gegebenenfalls zu überbrücken. In [37] wurde gezeigt, dass mehrsprachiges Information Retrieval durch den Einsatz von Metadaten verbessert werden kann: Petras wendete das EVM für die mit Thesaurustermen indexierte Fachdatenbank GIRT (German Indexing und Retrieval Testdatabase, vgl. [28]) an.

3.2. Konzeption eines Entry Vocabulary Moduls

In [9] wird detailliert der Bedarf von Ansätzen beschrieben, die Nutzern den Zugang zu Datenbasen mit unbekannten Metadaten erleichtern sollen. Die Autoren nennen mehrere Gründe hierfür: Datenbestände würden sowohl durch manuelle, als auch immer neuere und leistungsfähigere automatische Methoden um Metadaten ergänzt und durch sie erschlossen. Gleichzeitig ist die Nutzung dieses Mehrwerts einer Datenbasis nur dann möglich, wenn der Nutzer Kenntnis von der Verfügbarkeit, Struktur und Anwendungsmöglichkeit dieser Metadaten hat.

> „Indeed, the more that has been invested in the enhancement of the source, the richer the metadata and the more important this personal experience and familiarity become, and the less they can be used effectively or efficiently except by searchers who are familiar with them.“ (vgl. [9])

Die Autoren fügen an, dass in einem weiteren Trend immer mehr Datensammlungen untereinander vernetzt oder über das Internet zugänglich gemacht werden. Somit wird der Zugriff auf eine Fülle von verschiedenen Datengrundlagen gegeben. Je mehr die Anzahl der verschiedenen Metadatenstrukturen steigt, desto unwahrscheinlicher wird es, dass diese Masse an Daten und Wissen effektiv genutzt werden kann.

Vor diesem Hintergedanken, und ebenfalls vor dem Grundsatz des Information Retrieval, dem Nutzer auf der Suche nach Informationen eine wertvolle Unterstützung zu sein, lässt sich einfach ableiten, dass ein System benötigt wird, das dem Nutzer bei dem Umgang mit einer noch unbekannten Datenbasis die Unterstützung gibt, effektive, d.h. auf das kontrollierte Vokabular der Datenbasis angepasste Anfragen zu formulieren.

Das Konzept eines Entry Vocabulary Moduls wurde unter anderem auch in [17] vorgestellt. In dieser Arbeit wurden die vier zentralen Komponenten wie folgt bezeichnet:

- eine ausreichend große Datengrundlage zum Trainieren des Entry Vocabulary Index
- ein Part-of-Speech-Tagger, der Substantive aus Dokumenttexten extrahiert
- ein Algorithmus, der die Beziehung zweier Begriffe anhand der Wahrscheinlichkeit ihrer Kooexistenz in einem Dokument errechnet
- das grundlegende Retrievalsystem, das die Suchanfrage entgegennimmt und die Ergebnisse auflistet

Die Autoren von [17] verwenden zur Berechnung der Wahrscheinlichkeit der Zusammengehörigkeit eines Terms und einer Klassifikationsangabe die folgende Formel nach Dunning in [11]:

$$W(C,t) = 2\left[logL(p_1, a, a+b) + logL(p_2, c, c+d) - logL(p, a, a, +b) - logL(p, c, c+d)\right]$$

$$\text{where } logL(p,n,k) = klog(p) + (n-k)log(1-p)$$

$$\text{and } p_1 = \frac{a}{a+b},\ p_2 = \frac{c}{c+d},\ \text{and } p = \frac{a+c}{a+b+c+d}$$

> „[...] for each pair of word/phrase terms t and classifications C [...] where a ist the number of document titles/abstracts containing the word or phrase and classified by the classification; b is the number of document titles/abstracts containing the word or phrase but not the classified by the classification, c is the number of titles/abstracts not containing the word or phrase but is classified by the classification; and d is the number of document titles/abstracts neither containing the word or phrase nor being classified by the classification.“ (vgl. [17])

Vergleichbar mit den Systemen, die durch Gey und Buckland vorgestellt wurden, haben Schatz/Johnson/Cochrane in [40] ein System vorgestellt, dass ebenfalls Terme des kontrollierten Vokabulars vorschlägt. Hierbei werden sowohl Thesauri als auch Auflistungen von koexistenten Begriffen in Dokumenten verwendet. Die Autoren fassen eine zentrale Motivation ihrer Arbeit wie folgt zusammen:

> „By providing easy access to these various term suggestion mechanisms, we hope to encourage searchers to use them before attempting to access bibliographic records. This would reverse the current state of bibliographic as well as full-text retrieval, in which thesauri and other means of term suggestion (assuming they are even available) are typically accessed only after an initial bibliographic query yields either too few or too many hits.“ (vgl.[40])

Dieser Ansatz verdeutlicht den Nutzen von Entry Vocabulary Modulen: Während der Transfer des Vokabulars der eigenen Suchanfrage auf das Vokabular der Datenbasis in der klassischen Form des Retrieval durch den Nutzer (bspw. durch die Anwendung von Kategorisierungsschemata oder Schlagwortverzeichnissen) erbracht werden muss, kann ein System, das in der Lage wäre, das Vokabular der Anfrage anzupassen, dem Nutzer diese Arbeit erleichtern (interaktive Lösung) oder abnehmen (automatische Lösung). Im Falle von Referenzdatenbanken wird mit dem Ansatz der Entry Vocabulary Module versucht, die Leistung der Metadatenerschließung für einen Nutzer ohne Vorkenntnisse überhaupt nutzbar und für einen Nutzer mit Erfahrung mit dem System einfacher nutzbar zu machen.

Bei dem Datenbestand des Fachinformationsverbundes handelt es sich, wie bereits beschrieben, ebenfalls um eine durch Metadaten und eine Klassifikation erschlossene Datensammlung. Es ist zu vermuten, dass die Retrievalergebnisse eines nicht mit der Datenbasis vertrauten Nutzers durch den Einsatz eines Entry Vocabulary Moduls und die damit verbundene Aktivierung der Metadaten gesteigert werden kann. Allerdings ist hierbei zu beachten, dass die Metadaten des Datenbestands des FIV nicht

im ähnlichen Maße kryptisch sind wie beispielsweise die Metadaten der von Gey vorgestellten Bestände. Es handelt sich viel mehr um eine Sammlung von Begriffen, die aus Ländernamen, gebräuchlichen Begriffen und Themenangaben besteht. Dadurch ist möglicherweise nicht die gleiche Leistungssteigerung zu erwarten, die ein System erzielen könnte, das eine Anfrage zuverlässig um gänzlich kryptische Angaben erweitert. Es ist möglich, dass die Anfragen von sich aus schon einen Anteil an kontrollierten Vokabular beinhalten. Dies kann ein weiterer Ansatz für Untersuchungen sein.

Der grundlegenden Vermutung folgend, dass sich ein EVM positiv auf die Retrievalleistung auswirkt, wird im nächsten Kapitel die Entwicklung eines solchen Systems für den vorliegenden Datenbestand beschrieben und später evaluiert.

Teil III.

Entwicklung

4. Rahmenbedingungen der Entwicklung

Nachdem im Rahmen des zweiten Teils der Arbeit sowohl der Kooperationspartner vorgestellt wurde und die grundlegenden Information Retrieval Modelle sowie die Vokabularproblematik vorgestellt wurden, beginnt mit diesem Teil die Dokumentation der Entwicklung des Systems.

Im Vorfeld der Entwicklung des Retrieval Systems sollen im Folgenden die gegebenen Voraussetzungen untersucht und ein Überblick über die Ansätze zur Entwicklung eines Entry Vocabulary Moduls gegeben werden. Im Rahmen dieser Bestandsaufnahme wird zunächst in Abschnitt 4.1 die Datengrundlage des FIV untersucht und in später in Abschnitt 4.2 ab Seite 32 verschiedene Anforderungen und Nutzungssituationen beschrieben.

4.1. Analyse der Datengrundlage

Im Rahmen der Vorbereitungen des Indexierens steht eine Analyse der Datengrundlage im Vordergrund.

Entscheidend für die Wahl des Retrievalsystems ist zu vorderst das Medienformat der zu durchsuchenden Informationen: Handelt es sich um Text, Bild, Ton, Video, Hypermedia oder eine sonstige Form von Datenbestand?

Wichtige Fragen sind zunächst bezüglich der grundlegenden Architektur des Datenbestands zu klären: Handelt es sich um eine Datengrundlage, die aus einer monolithischen Sammlung oder aus mehreren Fragmenten zusammengesetzt ist? Haben die einzelnen Dokumente mehrere Datenfelder oder handelt es sich um reinen Volltext? Falls es sich um einen Datenbestand mit mehreren Feldern handelt: Welche Felder verfügen über Freitext, welche über kontrolliertes Vokabular? Handelt es sich möglicherweise um eine Sammlung von Hypertextdokumenten mit einer Verknüpfungsstruktur? Liegt die Datengrundlage gar in einem proprietären Format vor, das nicht ohne weiteres indexiert werden kann? Diese und weitere Fragen müssen zunächst geklärt werden, um eine geeignete Retrievalstrategie zu wählen und das Indexieren zu ermöglichen.

4.1.1. Art der verzeichneten Informationen

Die Datengrundlage des Fachinformationsverbundes ist, wie in Abschnitt 1.4.2 erwähnt, eine thematisch spezialisierte Form der Referenzdatenbank. Entsprechend ergeben sich andere Bedingungen für und Anforderungen an ein Information Retrieval System als bei einer Datengrundlage, die etwa ganz, oder zu großen Teilen aus Volltexten bestehen.

Während im Retrievalprozess von Volltexten die frei formulierte Anfrage mit dem freien Vokabular der Volltexte abgeglichen werden muss, ergibt sich in der vorliegenden Form einer Referenzdatenbank die Situation, dass die verzeichneten Objekte durch mehrere zusätzliche Attribute (Titel, Autor, Zusammenfassung, Datum, Deskriptoren und weitere) beschrieben werden. Diese Attribute unterscheiden sich in Umfang, Vokabular und Datentyp deutlich voneinander: Das Titel-Attribut ist ein kurzes, frei formuliertes Textfeld, während die Datumsangabe einen, einem gewissen Format folgenden, quantitativen Inhalt hat. Die Zusammenfassung ist ein umfangreicheres, frei formuliertes Textfeld und eine Themenbeschreibung oder Klassifikation ein kurzes, für kontrolliertes Vokabular genutztes Textfeld. Es ist also sinnvoll, in der Retrievalstrategie einen Ansatz zu wählen, der die Heterogenität der verschiedenen Attribute und die Suche auf die verschiedenen Indexfelder berücksichtigt.

Da im Falle dieser Referenzdatenbank viele Informationen über den Inhalt eines verzeichneten Dokuments in den Metadaten gespeichert sind, wird die Hauptaufgabe des Indexierens sein, diese Metadaten zusätzlich zu den grundlegenden Informationen wie Titel und Zusammenfassung feldgerecht zu indexieren und so entsprechend durch Anfragen exklusiv oder inklusiv nutzbar zu machen. Die Umsetzung des Einlesens und der Indexierung findet sich in Abschnitt 7.

4.1.2. Datenformat der Datengrundlage

Die Datengrundlage des Fachinformationsverbundes für Internationale Beziehungen und Länderkunde besteht in der vorliegenden Version aus 600 XML-Dateien, die jeweils zwischen circa 3 und 7,5 MB groß sind und somit insgesamt rund 3,4 GB an Speicherplatz belegen. Eine Datei verzeichnet damit zwischen drei und siebeneinhalb Millionen Zeichen. In jeder, bis auf die letzte der 600 Dateien sind jeweils 1000 Einträge und damit insgesamt 599.723 Dokumente verzeichnet. Entsprechend beschreibt ein verzeichnetes Dokument im Schnitt 5,7 KB.

Die Komplexität der Architektur des XML-Auszugs der Datenbank des Fachinformationsverbundes ist gerade im Vergleich zur ebenfalls zu Testzwecken verwendeten Datenbasis GIRT sehr hoch. Eine beispielhafte, einfache Analyse einer FIV-Datendatei

(swp_fiv_25.xml) zeigt, dass von insgesamt 4.744.730 Zeichen nur 971.715 Zeichen (rund 20%) in Form von Nutzdaten gespeichert sind. Der Rest der Zeichen entfällt auf den XML-Steuersatz. Ein weiteres Beispiel für die Komplexität der XML-Architektur ist das 929 Zeilen umfassende XML-Schema, dass alle Funktionen und Datenfelder des Datenbankauszugs beschreibt. Darüber hinaus wird ein Großteil von Informationen in Elementattributen gespeichert.

Im Vergleich zu diesem Datensatz verzeichnet GIRT in Version 3 insgesamt 76.128 Dokumente in einer XML-Datei mit einem Umfang von 157 MB. Dies ergibt einen Speicherplatzumfang von 2,1 KB pro Dokument, wobei besonders zu beachten ist, dass entsprechend der Analyse in Abschnitt 10.2.1 auf Seite 84 nahezu alle GIRT-Dokumente über speicherintensive Zusammenfassungen verfügen, die in FIV verzeichneten Dokumente aber nur zu gut einem Viertel mit Zusammenfassungen ausgestattet sind. Entsprechend ist der Anteil von XML-Overhead bei GIRT3 deutlich kleiner als bei der FIV-Datenbasis, der Nutzdatenanteil von GIRT liegt bei rund 67% (hochgerechnet).

Zusammenfassend lässt sich sagen, dass die Datengrundlage des FIV eine sehr hoch entwickelte und komplexe Architektur hat. Dies führt entsprechend zu einem ebenfalls komplizierteren Ansatz bei dem Parsing der Dateien, was im Rahmen der Indexierung notwendig werden wird.

4.1.3. Thematische und geopolitische Abdeckung

Die Datenbank des FIV verzeichnet Dokumente, die sich zum Großteil mit den Themengebieten Staat- und Gesellschaft, nationale und internationale Wirtschaft, internationale Politik und Sicherheit befassen. Geopolitisch beziehen sich mehr als die Hälfte der verzeichneten Dokumente auf Europa, europäische Organisationen und die in der NATO kooperierenden Länder. Weitere wichtige geographische Regionen schließen Afrika und den Nahen Osten, Nord- und Südamerika und Asien und Ozeanien neben anderen mit ein. [45]

Die Datengrundlage verzeichnet zu 65% Bücher und wissenschaftliche Publikationen, zu 25% monographische Veröffentlichungen und zu jeweils 5% Periodika und Jahrbücher sowie amtliche Veröffentlichungen.

4.1.4. Klassifikation und Thesaurus

Die Datengrundlage des Fachinformationsverbundes zeichnet sich durch eine weitreichende Erschließung durch Metadaten aus. Die verzeichneten Dokumente werden sowohl anhand von Sach- und Regionalklassen kategorisiert, als auch durch die Vergabe

von themenbezogenen Schlagworten inhaltlich beschrieben. Mehr als 98% der verzeichneten Dokumente sind durch einen oder mehrere Deskriptoren beschrieben, im Durchschnitt über alle verzeichneten und entsprechend erschlossenen Dokumente verfügt jedes Dokument über rund 12 Deskriptoren. 85% der Dokumente sind außerdem über die Klassifikation kategorisiert (vgl. [43] und Tabelle 4.1). Neben den gröberen Schlagworten werden außerdem feinere Beschreibungen durch Aspektdeskriptoren zugewiesen. Darüber hinaus wird in rund einem Viertel der Dokumente eine Zusammenfassung des verzeichneten Dokuments ergänzt. Einen Überblick über die Klassifikation bieten die Übersichten in [14] und [13] und über den Thesaurus die Auflistung in [12].

	Anzahl	Anteil an gesamt
Dokumente insgesamt	599723	100,0%
Dokumente mit Titel	599973	100,0%
Dokumente mit thematischen Deskriptoren	588747	98,2%
Dokumente mit geopolitischen Deskriptoren	517164	86,2%
Dokumente mit Klassifikationsangaben	513317	85,6%
Dokumente mit Zusammenfassungen	156933	26,2%
	Anzahl gesamt	Anzahl pro erschl. Dokument
Verzeichnete thematische Deskriptoren	7144208	12,1
Verzeichnete geopolitische Deskriptoren	633636	1,2
Verzeichnete Klassifikationsangaben	617567	1,2

Tabelle 4.1.: Erschließung der Datenbasis des FIV durch Metadaten

4.1.5. Mehrsprachigkeit

Die in der Datenbasis des Fachinformationsverbundes verzeichneten Dokumente sind zu einem Großteil nicht auf Deutsch verfasst. Sprachlich dominieren die englischen Dokumente mit 52% die Datengrundlage. Deutsche Dokumente machen 26% des Umfangs aus, französische rund 12% und spanische 4%, während der Rest der Dokumente in anderen Sprachen verfasst ist.

Da es sich bei der Datengrundlage um ein Verzeichnis und nicht um eine Volltext-Sammlung handelt, ist der Faktor der Mehrsprachigkeit eingegrenzt: In der vorliegenden Variante der Datengrundlage sind alle Dokumente mit deutschen Deskriptoren und auf Deutsch formulierten Klassifikationen ausgezeichnet. Die Zusammenfassungen der Dokumente sind, wo vorhanden, für deutsche Dokumente auf Deutsch und für fremdsprachliche Dokumente in der entsprechenden Fremdsprache oder auf Deutsch bzw. Englisch verfasst. Die gesamte Verteilung der wichtigsten Sprachen und der Anteil der fremdsprachlichen Dokumente mit Zusammenfassungen lässt sich aus Tabelle 4.2 entnehmen. Die Werte dieser Tabelle sind das Ergebnis einer eigens im Vorfeld durchgeführten Analyse.

	Anzahl Dokumente	Anzahl Abstracts	Dok. mit Abstracs	Anteil Dok. an ges.
Englisch	311036	88240	28,4%	51,9%
Deutsch	158970	35041	22,0%	26,5%
Französisch	73116	13219	18,1%	12,2%
Spanisch	25910	10046	38,8%	4,3%
Russisch	9918	4525	45,6%	1,6%
Arabisch	6704	2682	40,0%	1,1%
Portugiesisch	3784	1583	41,8%	0,6%
Italienisch	1236	313	25,3%	0,2%
Summe	590674	155649	26,4%	98,5%

Tabelle 4.2.: Sprachverteilung in der Datenbasis des Fachinformationsverbundes

Die Identifikation der Sprache des Dokuments ist nur über die eindeutige Auszeichnung je Objekt möglich. Die Sprache des beschriebenen Dokuments lässt dabei nur vage auf die Sprache der Zusammenfassung schließen, beispielsweise werden viele arabische Dokumente auf Englisch oder Deutsch beschrieben, während die Titel entweder auf Arabisch, Französisch, Englisch oder Deutsch verfasst sind. Die Deskriptoren sind bis auf mehrere Namen von Organisationen in Deutsch verfasst.

Aufgrund der umfassenden Erschließung der Datengrundlage durch Deskriptoren und eine Klassifikation, die über alle Dokumente in einer Sprache verfasst sind, kann bei einer intensiven Nutzung des kontrollierten Vokabulars des Deskriptoren-Thesaurus und der Klassifikation von einer inhärenten Mehrsprachigkeit des Retrievalsystems ausgegangen werden. Über die Metadaten werden die Dokumente per Deskriptoren in einem einheitlichen, „supranatürlichsprachlichen“ Vokabular beschrieben. Dieser Ansatz scheint besonders geeignet, da nicht nur der gesamte Datenbestand, sondern selbst viele Dokumente in sich, wie beschrieben, sprachlich über mehrere Felder heterogen sind und die Metadaten eine gemeinsprachliche Basis bieten.

4.2. Analyse der Nutzersituation

Im Rahmen der Entwicklung eines IR-Systems ist es immer von zentraler Bedeutung, die endgültige Nutzung des Systems im Vorfeld der Entwicklung zu berücksichtigen und ggf. in die Entwicklung einfließen zu lassen. In dem Vorabgespräch mit dem Leiter des Fachinformationsbereichs der Stiftung Wissenschaft und Politik wurde unter anderem die übliche Nutzung der Datenbank durch professionelle Nutzer beschrieben, die sowohl in Abschnitt 4.2.1 berücksichtigt wird als auch später in 4.2.3 zum Tragen kommt. In Abschnitt 4.2.2 wird dagegen die Situation eines Nutzers mit semiprofessionellen Kenntnissen und Ansprüchen im Umgang mit einer Referenzdatenbank wie der des FIV erläutert. Abschnitt 4.2.3 fasst beide Situationen zusammen und stellt sie einander gegenüber.

4.2.1. Professionelle Nutzung

Die Datenbank des Fachinformationsverbundes Internationale Beziehungen wird von mehreren Benutzergruppen entwickelt und genutzt. Aufgrund der Struktur des Verbundes arbeiten viele verschiedene Gesellschaften wie Forschungsinstitute und Universitäten an der Datengrundlage. Alle diese Institutionen haben einen gemeinsamen wissenschaftlichen Hintergrund, entsprechend teilen ihre Mitarbeiter ein fundiertes Wissen bezüglich der Materie der internationalen Politik. Da es, wie in allen wissenschaftlichen Fachrichtungen, auch in dieser Disziplin eine spezialisierte Ausprägung der Sprache und des Vokabulars gibt, ist die zentrale Voraussetzung für die erfolgreiche Nutzung des Systems gelegt.

Zu der erfüllten Voraussetzung der Beherrschung des Vokabulars kommt hinzu, dass viele der angeschlossenen Institute separate Referate unterhalten, die sich hauptsächlich mit dem Information Engineering bzw. mit der Recherche in dieser Datenbank befassen. Da sich solche Nutzer intensiv mit dieser Datenbank auseinandersetzen, sie bisweilen selbst mitentwickeln, liegt ein tiefgehendes Verständnis für die Struktur, die Bezeichnung der Dokumentattribute, die Bezeichnungen der Klassifikation und das Retrievalsystem an sich, nahe. Dies führt zu einer Nutzung, die einer Datenbankabfrage nicht unähnlich ist. Beispielhaft kann die durch einen Fachreferenten des Fachinformationsbereichs der SWP formulierte Anfrage

„(5210=DG=Russland (seit 1991/92) OR
5210=DG=Rußland (Sowjetrepublik) OR
5210=DG=Rußland (vor 1917)) AND
DT=Rechtsextremismus“ (vgl. [7])

genannt werden. In solchen, äußerst präzisen Anfragen werden explizite Deskriptoren in bestimmten Attributfeldern gesucht, verknüpft durch Boolesche Logik. Eine Anfrage in dieser Form nutzt die architektonischen Fähigkeiten der Datengrundlage, ist aber auch durch die eindeutige Charakteristik einer Booleschen Anfrage durch die in Abschnitt 2.1.1 ab Seite 10 beschriebenen Vor- und Nachteile geprägt: Beispielsweise können auf diese Weise sehr große Dokumentlisten generiert werden, die zusätzlich ausgewertet werden müssen.

4.2.2. Semiprofessionelle Nutzung

Durch mehrere Onlineportale wird die Datenbank des Fachinformationsverbundes Internationale Beziehungen und Länderkunde auch semiprofessionellen Anwendern oder Laien angeboten.

Selbst für einen auf dem Feld der internationalen Politik erfahrenen Nutzer, nehmen wir als Beispiel einen Mitarbeiter aus einem Institut, kann der erfolgreiche Einsatz eines Retrievalsystems für die Datenbasis wegen des hohen Anteils an in kontrolliertem Vokabular verschlüsselten Informationen und der großen Anzahl an verschiedenen Metainformationen in verschiedenen, namentlich unbekannten Attributfeldern problematisch werden. Während ein in politischer Wissenschaft erfahrener Nutzer die richtigen Schlagworte für eine Anfrage formulieren kann, bedeutet das noch nicht, dass das System ohne weiteres seine Anfrageterme neben den freitextlichen Feldern wie der Zusammenfassung und dem Titel auch mit den kontrollierten Feldern der Deskriptoren oder Klassifikationen vergleicht. Der erfahrene, semiprofessionelle Nutzer würde also eine Anfrage in natürlicher Sprache oder mit Schlagworten formulieren, aber nicht ohne zusätzliche Einweisung auch die in einem Verzeichnis hilfreichen Deskriptoren- und Klassifikationsattribute in den entsprechenden Indexfeldern konkret ansprechen können.

Im Fall eines auch fachlich eher unerfahrenen Nutzers, beispielsweise eines Studenten der Politikwissenschaften am Studienbeginn, kommt gegebenenfalls erschwerend hinzu, dass der Mangel an Fachvokabular dazu führt, dass es nahezu unmöglich wird, eine für eine Datenbank, die zu großen Teilen aus einer kontrollierten Form und damit nur einem Ausschnitt des Fachvokabulars besteht, sinnvolle Anfrage zu formulieren. Weder Hintergrundinformationen über die Struktur der verzeichneten Dokumente noch umfassendes Vokabularwissen stehen diesem Nutzer zur Verfügung.

4.2.3. Unterschiedliche Anwendung und Anforderungen

Es gibt einen zentralen Unterschied in den Anforderungen der professionellen und der semiprofessionellen Nutzer. Für die professionellen Nutzer liegt der Fokus der Leistungsfähigkeit des Systems auf dem Recall, also der Leistungsfähigkeit zum Auffinden aller relevanten Dokumente. Der professionelle Nutzer wird üblicherweise durch einen Rechercheauftrag dazu veranlasst, dass System zu durchsuchen. Ein solcher Auftrag erfordert, dass der die Recherche durchführende Nutzer eine umfassende Zusammenstellung von Literatur erarbeitet. Entsprechend wird vorausgesetzt, dass der professionelle Nutzer mitunter mehrere hundert aufgelistete Dokumente bewertet und im Fall der Relevanz vermerkt. Hier steht im Mittelpunkt, dass durch eine extensive Suche möglichst viele Informationen zusammengetragen werden.

Die semiprofessionelle Suche ergibt sich aus einem anderen Kontext. Üblicherweise wird aus individueller Initiative die Datenbasis durchsucht, um mehrere Informationen, oder im Falle des FIVs, mehrere Literaturhinweise zu einem bestimmten Thema zu finden. In diesem Rahmen ist es selten notwendig, einen absolut erschöpfenden

Überblick über alle in der Datenbank verzeichneten, relevanten Dokumente zu erhalten und dafür ggf. in Kauf zu nehmen, dass mehrere hundert Einträge einer Auflistung zu bewerten sind. Entsprechend ist für den semiprofessionellen Nutzer von zentralem Interesse, dass die relevanten Dokumente besonders hoch auf der Ergebnisliste des IR-Systems aufgeführt werden: Eine hohe Precision, also die Rate von relevanten zu nicht relevanten Dokumenten, ist damit üblicherweise wichtiger.

Für beide Benutzergruppen kann allerdings der Einsatz eines Entry Vocabulary Moduls sinnvoll sein. Einerseits profitiert der semiprofessionelle Nutzer durch die Nutzung der Vorteile der aufwändigen Datenarchitektur und -erschließung. Damit verbessert sich die Ausgangssituation dafür, das System zielführend zu nutzen und so die Retrievalqualität der Anfragen zu steigern. Um den Aufwand, das System zu nutzen, so gering wie möglich zu halten, würde sich für den semiprofessionellen Nutzer empfehlen, den EVM-Prozess im Hintergrund der Suchanfrage automatisch abzuwickeln. Andererseits profitiert der professionelle Nutzer davon, dass er nicht gezwungen ist, in dem umfassenden Thesaurus von Deskriptoren und der Klassifikation zunächst die passenden Begriffe zu suchen, sondern dass ihm vielmehr im Rahmen einer interaktiven Lösung die passenden Begriffe zur Erweiterung der Anfrage angezeigt werden. Im Rahmen der professionellen Nutzung bietet sich eine interaktive Lösung an, da durch eine auf diese Weise verbesserte Anfrage im nachhinein, beim Auswerten der gefundenen Dokumente, viel Zeil sparen kann. Der interaktive Prozess der Auswahl von Deskriptoren durch einen erfahrenen Nutzer wird potentiell erfolgreicher sein als die automatische Ergänzung einer Anfrage durch das System selbst. Im Folgenden wird, der natürlichsprachlichen Art der gegebenen Evaluierungsanfragen (vgl. die Auflistung der gegebenen Evaluierungsanfragen im Anhang ab Seite 108) Rechnung tragend, ein System zur automatischen Erweiterung der Suchanfrage entwickelt, da zu erwarten ist, dass natürlichsprachliche Anfragen eher von einem semiprofessionellen Nutzer als von einem professionellen Nutzer formuliert würden.

5. Das dynamische Entry Vocabulary Modul

5.1. Ansatz

Das von Gey in [17] vorgestellte System ist ein globaler (d.h. auf den gesamten Datenbestand bezogener) Ansatz zur Konstruktion eines Entry Vocabulary Index, in dem Terme des freien Vokabulars mit den kontrollierten Vokabeln der Metadaten auf Basis von probablistischen Untersuchungen verknüpft werden. Diese Verknüpfung basiert auf einer statistischen Analyse der Koexistenz von Termen und vergebenen Metainformationen für ein gegebenes Dokument.

Diese Entwicklung eines zusätzlichen Datenkonstrukts, des Entry Vocabulary Index, ist dabei grundsätzlich nicht zwingend erforderlich. Eine dynamische, lokale (d.h. auf eine Gruppe von potentiell relevanten Dokumenten angewendete) Lösung minimiert den Aufwand der Pflege und der stetigen Aktualisierung eines zusätzlichen Datenbestands. Da die Datenbasis des FIV stetig wächst und aktuellere Themen, mit denen sich die Beiträge befassen, ebenfalls in vielen Fällen neues, freies Vokabular mit sich bringen, wäre eine regelmäßige Neuberechnung notwendig. Schließlich ist zu erwarten, dass ein großes Interesse daran besteht, auch die neuesten Ergänzungen der Datenbank effektiv aufzufinden. Darüber hinaus wird in [50] beschrieben, dass zumindest für Terme aus dem Freitext eines Datenbestands ein lokaler, dem Relevance Feedback verwandter Ansatz nicht schlechter geeignet sein muss als eine globale Berechnung.

Ein weiteres Argument gegen eine globale Auswertung des paarweisen Auftretens von Termen des freien Vokabulars und Deskriptoren ist die begrenzte Anzahl von Dokumenten mit Freitexten in Form von Zusammenfassungen (rund ein Viertel aller Dokumente). Eine entsprechende Auswertung würde die Deskriptoren der mit Zusammenfassungen ausgestatteten Dokumente voraussichtlich anders in Relation zu den Termen des Freitextes stellen als es bei den Dokumenten der Fall wäre, in denen die Deskriptoren nur mit den Termen der Titel in Relation gebracht werden könnten. Würde sich eine solche globale Analyse nur auf die Zusammenfassungen beziehen, könnte nur ein Viertel der Dokumente entsprechend ausgewertet werden. Bei einer dy-

namischen Lösung können dagegen auch Metadaten zueinander in Relation gebracht werden, indem in einem Suchprozess anhand der bereits extrahierten Metadaten gesucht wird und weitere Deskriptoren extrahiert werden können, die besonders häufig in den gefundenen Dokumenten auftreten. Ein solcher Ansatz würde bei dem Datenbestand des FIV so gut wie alle Dokumente in einer Auswertung mit einbeziehen, da nahezu alle Dokumente mit Deskriptoren erschlossen sind.

5.2. Prozessablauf

Dem eingesetzten dynamischen Entry Vocabulary Modul wird im Suchprozess zunächst die ursprüngliche Anfrage übergeben. Diese Anfrage wird zuerst in einem Suchdurchlauf auf die Indexfelder des freien Vokabulars, also die Felder der Titel und Zusammenfassungen gerichtet. Nach diesem Suchvorgang werden die Deskriptoren einer gegebenen Anzahl der am höchsten bewerteten Dokumente extrahiert, mit den Dokumentbewertungen verknüpft und statistisch, wie in Abschnitt 5.3 beschrieben, ausgewertet. Diese Deskriptoren können nun eingesetzt werden, um die Anfrage zu erweitern. Darüber hinaus können, wie bereits erwähnt, die Deskriptoren erneut dazu eingesetzt werden, in einem weiteren Suchlauf passende Deskriptoren aus anderen Feldern des Index mit kontrolliertem Vokabular zu extrahieren. Durch die einzelnen Vorgänge, die teilweise auf den Ergebnissen der zuvor extrahierten Deskriptoren basieren, ergibt sich eine kaskadenförmige Struktur: Der eigentliche Transfer von freiem Vokabular in das kontrollierte Vokabular passiert an der Spitze der Kaskade zuerst, alle weiteren Elemente der Kaskade versuchen, in Assoziation stehende Deskriptoren zu finden und, falls sie besonders häufig auftreten, der Anfrage zu ergänzen. Hierbei ist es möglich, dass eine große Anzahl von Deskriptoren gefunden wird und unter anderem auch Mehrfachnennungen in der Gruppe der ergänzten Deskriptoren auftreten. Dies ist allerdings gewünscht: Eine Mehrfachnennung bedeutet, dass ein Deskriptor anscheinend besonders gut geeignet ist, da durch Assoziation belegt werden konnte, dass der Begriff sich thematisch im gleichen Feld bewegt, wie die Ausgangsfrage. Es ist zu beachten, dass durch die bei einer langen Kaskade große Anzahl von Deskriptoren durch Mehrfachnennungen große Gewichtungen auftreten können, die, ist ein die ursprüngliche Frage ergänzendes System das Ziel, das Gewicht der ursprünglichen Anfragepassage leicht übersteigen können. Ein dynamisches Entry Vocabulary Modul lässt sich an die Datengrundlage und die entsprechende Aufgabe sehr flexibel anzupassen. Einen Überblick über einige Umsetzungen, die mit einem solchen System realisiert werden können, gibt Abschnitt 5.4.

Deskriptor	Document Score
Europäische Union	0,99999994
Mittel- und osteuropäische Länder	0,99999994
Erweiterung von undBeitritt zu internationalem Akteur	0,99999994
Sicherheitspolitische Interessen	0,99999994
Wirkung/Auswirkung	0,99999994
Gemeinsame Außen- und Sicherheitspolitik (EU)	0,99999994
Europäische Sicherheits-und Verteidigungspolitik (EU)	0,99999994
Regionale internationale Sicherheitsstruktur	0,99999994
Sicherheitspolitische Zusammenarbeit	0,99999994
Russische Föderation	0,99999994
North Atlantic Treaty Organization	0,99999994
Regionaler internationaler Konflikt	0,99999994
Balkan	0,99999994
Türkei	0,99999994
Zypern	0,99999994
Demokratisierung	0,99999994
Rechtsordnung	0,99999994
Status und Rolle im internationalen System	0,8164988
Reform	0,8164988
Entscheidungsverfahren bei internationalem Akteur	0,8164988
Zuständigkeit	0,8164988
Wirtschaftliche Entwicklung	0,8164988
Entwicklung internationalen Akteurs	0,8164988
⋮	⋮

Tabelle 5.1.: Erster Schritt: Extraktion der Metadaten und Verknüpfung mit dem Score des jeweiligen Dokuments

5.3. Prozess der statistischen Auswertung

Die statistische Auswertung geschieht in dem vorgestellten System auf der Basis der Verknüpfung des durch das Retrieval System vergebenen Scores eines gefundenen Dokuments, mit den in diesem Dokument verzeichneten Deskriptoren (vgl. Tabelle 5.1 als Bespiel für die Anfrage `+eu +erweiterung +osteuropa`). Durch Mehrfachnennungen der selben Deskriptoren in mehreren verschiedenen Dokumenten addieren sich die einzelnen Scores auf (vgl. Tabelle 5.2). Formal werden also der Scoring-Algorithmus von Lucene (vgl. [22], Seite 78) mit einer normalisierten Addition kombiniert, entsprechend:

$$documentscore = \sum_{t \in q} tf(t \in d) \times idf(t) \times boost(t.field \in d) \times lengthNorm(t.field \in d)$$

$$descriptorscore = \frac{\sum_{d \in Ret} documentscore_{d_{t_{raw}}}}{documentscore_{max_{raw}}}$$

Deskriptor	Kumulierter Score	Normalisierter Score
Europäische Union	8.079417	1.0
Mittel- und osteuropäische Länder	7.253403	0,8977632
Erweiterung von und Beitritt zu internationalem Akteur	5.080429	0,6288113
Internationale regionale politische Integration	4.358143	0,53941303
Wirkung/Auswirkung	3.6277335	0,4490093
Erweiterung von undBeitritt zu internationalem Akteur	2.9897354	0,37004346
EU/EG sowie	0,99999994	0,12345391
Mittel- und osteuropäische Länder	0,99999994	0,12345391
Gemeinschaft Unabhängiger Staaten	0,7355408	0,09110038
Frankreich	0,63799816	0,07897017
Russische Föderation	0,58843267	0,07278129
Weißrußland	0,5148786	0,063622
⋮	⋮	⋮

Tabelle 5.2.: Zweiter Schritt: Akkumulierung der Mehrfachnennungen, Normalisierung der Scores

Passage, die der Suchanfrage hinzugefügt werden kann:
subject:(Europäische Union)^1.0 subject:(Mittel- und osteuropäische Länder)^0.8977632 subject:(Erweiterung von und Beitritt zu internationalem Akteur)^0.6288113 subject:(Internationale regionale politische Integration)^0.53941303 subject:(Wirkung/Auswirkung)^0.4490093 subject:(Erweiterung von undBeitritt zu internationalem Akteur)^0.37004346 geo:(EU/EG sowie Mitgliedsländer)^1.0 geo:(Mittel- und osteuropäische Länder)^0.6627917

Tabelle 5.3.: Dritter Schritt: Formulierung der Anfragenpassage, die der ursprünglichen Anfrage hinzugefügt werden kann

Die nach dieser Auswertung potentiell relevanten (d.h. häufigsten und/oder am höchsten bewerteten) Terme können zur Anfragenerweiterung hinzugezogen werden. Bei einer Analyse der 100 am höchsten bewerteten Dokumente auf ihre thematischen Deskriptoren werden bei durchschnittlich 12,2 thematischen Deskriptoren pro Dokument insgesamt rund 1220 Deskriptoren-Nennungen ausgewertet und auf eine kleine Anzahl der Deskriptoren reduziert, die die höchsten Gesamtergebnisse erzielen konnten.

Zwei Methoden zur Limitierung der Anzahl der hinzugefügten Deskriptoren sorgen dafür, dass nur die am besten geeigneten Metadaten ergänzt werden: Ein Schwellenwert verhindert, dass viele Deskriptoren mit einem zu geringen Wert ergänzt werden und eine festgelegte, maximale Anzahl von Deskriptoren verhindert, dass die Query um zu viele Terme erweitert wird. Da der Typ (geopolitisch/thematisch/etc.) jedes einzelnen Deskriptors eindeutig ist, kann in der Suchanfrage direkt mit dem Deskriptor auf das passende Indexfeld gesucht und ggf. auch der erzielte Score als Gewicht eingesetzt werden (vgl. Tabelle 5.3 und Abschnitt 6.1 für eine Erklärung der Syntax).

5.4. Variabilität der Umsetzung

Ein System zur dynamischen Extraktion und Ergänzung von Metainformationen hat eine Vielzahl von Anpassungsmöglichkeiten. In den folgenden drei Unterkapiteln sollen die (teilweise bereits beispielhaft erwähnten) grundlegenden Möglichkeiten zur Modifikation genannt werden.

5.4.1. Termorientierte vs. anfragenorientierte Extraktion

Einerseits ist es möglich, die ursprüngliche Anfrage Term für Term in das kontrollierte Metadatenvokabular zu übersetzen, andererseits kann die gesamte Anfrage in einem Durchgang eingesetzt werden, um als Ergebnis Terme des kontrollierten Vokabulars zu erhalten, die den gesamten Kontext der Anfrage beschreiben, anstelle von nur jeweils einem Term.

Die Übersetzung der einzelnen Terme kann dazu verwendet werden, eine gesamte Anfrage Schritt für Schritt zu reformulieren und die ursprüngliche Anfrage so zu einer gänzlich auf Metadaten basierenden Anfrage mit Paaren aus Indexfeldnamen und geeignetem Deskriptor zu übersetzen. Hierbei ist zu beachten, dass eine geringe Anzahl von besonders gut geeigneten Deskriptoren extrahiert werden muss, um die ursprüngliche Anfrage nicht zu stark zu verwässern.

Die Übersetzung der gesamten Anfrage eignet sich potentiell am besten, um eine originalsprachliche Anfrage zu ergänzen, da durch die Auswertung der gefundenen Dokumente, welche für die gesamte ursprüngliche Anfrage als relevant bewertet wurden, eine Sammlung von Metadaten zusammengestellt werden kann, die den Sinn der gesamten Anfrage möglicherweise besser abbildet. Hierbei ist allerdings zu beachten, dass die Qualität der ursprünglichen Anfrage deutliche Auswirkungen auf die Qualität der ergänzten Deskriptoren hat. Die extrahierten Deskriptoren spiegeln dabei immer die häufigsten und relevantesten Inhalte der im Kontext der ursprünglichen Anfrage gefundenen Dokumente.

5.4.2. Eindimensionale vs. mehrdimensionale Ergänzung

In einem Datenbestand wie dem des FIV, in dem unterschiedliche Metadatenfelder mehrere verschiedene inhaltliche Bezugssysteme abdecken, ist es möglich, die ursprüngliche Anfrage nicht nur in ein Bezugssystem zu transferieren (beispielsweise den thematischen Bezug) sondern ggf. auch weitere Bezugssysteme (beispielsweise auch den geopolitischen Bezug oder die klassifikatorischen Angaben) zu berücksichtigen.

Grundsätzlich ist davon auszugehen, dass in einem Index mit Dokumenten, die über

mehrere Datenfelder mit Deskriptoren verfügen, die Deskriptoren mehrerer verschiedener Datenfelder ein Dokument besser beschreiben können, da es auf diese Art und Weise in mehreren Bezugssystemen beschrieben wird. Entsprechend ist zu erwarten, dass durch die Erweiterung der Suchanfrage um Metadaten aus mehreren Bezugssystemen bessere Ergebnisse erzielt werden können.

5.4.3. Einmalige vs. kaskadenförmige Extraktion

Die grundlegende Umsetzung dieses EVM basiert auf der Verwendung der ursprünglichen Anfrage und der folgenden Extraktion der Metadaten jener Dokumente, die für diese Anfrage relevant scheinen. Es ist allerdings auch möglich, die Ergebnisse der ersten Extraktion weiter zu verwenden und wiederum zur Extraktion von anderen Metainformationen zu nutzen. Hierbei werden die bereits extrahierten Deskriptoren auf alle Dokumente, die über ein solches Feld verfügen, angewandt und aus den relevantesten Dokumenten die Metadaten eines anderen Feldes extrahiert und statistisch bewertet.

Hierbei ist zu beachten, dass sich eine Kaskade mit mehreren aufeinander aufbauenden Elementen inhaltlich ggf. von dem ursprünglichen Informationsgehalt der Anfrage entfernen kann, da jede Metainformation, die nicht ideal geeignet ist, die Ergebnisse einer weiteren Verwendung verschlechtert.

Andererseits ist zu berücksichtigen, dass bei einem System, das durch einen begrenzten Anteil von Zusammenfassungen in der Auswertung limitiert wird, dadurch profitieren könnte, über Metadaten noch einen größeren Kreis von Dokumenten anzusprechen und auszuwerten.

6. Verwendete Software

Das im Folgenden beschriebene System wurde in Java programmiert, im Rahmen der Entwicklung wurden mehrere Open Source Bibliotheken verwendet. Fundament des Retrieval Systems ist die Klassenbibliothek Apache Lucene[1] in Version 1.9.1. Zum Parsen während der Indexierung der XML-Dateien wird Jakarta Commons Digester[2] in Version 1.7 verwendet. Im Rahmen des Suchprozess wird ebenfalls das Blind Relevance Feedback Modul der Retrieval-Systeme der Universität Hildesheim eingesetzt (vgl. beispielsweise den Einsatz in [24]). Zur Evaluierung des Systems kommen sowohl ein Relevanzbewertungsprogramm des Informationszentrum Sozialwissenschaften in Bonn als auch das an der Universität Hildesheim von Viola Barth und Joachim Pfister (vgl. [6]) nach Java portierte Programm Trec_Eval (im Original von Gerard Salton und Chris Buckley) in Version 0.7 zum Einsatz. Im Folgenden werden Lucene und Digester, die für das System zentralen Programme, kurz beschrieben.

6.1. Lucene

Otis Gospodnetic und Erik Hatcher geben in [22] einen umfassenden Einblick in Lucene. In ihrem Buch wird die Open Source Software wie folgt beschrieben:

> „Lucene is a high performance, scalable Information Retrieval library. [...] Lucene is a mature, free, open-source project implemented in Java; it's a member of the popular Apache Jakarta family of projects, licensed under the liberal Apache Software License. As such, Lucene is currently, and has been for a few years, the most popular free Java IR library.“ (vgl. [22], Seite 7)

Lucene ist, seit seiner ersten Veröffentlichung im Jahr 2000, zu einer wichtigen Basis für viele kommerzielle und Open Source Projekte[3] sowie wissenschaftliche Arbeiten

[1] vgl. auch http://lucene.apache.org/
[2] vgl. auch http://jakarta.apache.org/commons/digester/
[3] vgl. auch http://wiki.apache.org/jakarta-lucene/PoweredBy

im Bereich des textbasierten Information Retrieval geworden. Lucene kann aus Textbeständen, und damit auch aus allen Datenbeständen, die sich zu Text konvertieren lassen, Indices erstellen und diese Indices durchsuchen. Da es sich um eine Klassenbibliothek handelt, ist Lucene an sich keine fertige und ohne Ergänzungen funktionstüchtige Suchmaschine, Lucene ist vielmehr die Basis, auf der ein Retrievalsystem entwickelt werden kann.

Da es sich bei dem vorgestellten System um einen Ansatz zur Verbesserung der Retrievalqualität durch Ergänzung der Anfrage handelt, ist eine vorausgehende Untersuchung der Fähigkeiten zur Anfrageerweiterung in Lucene notwendig. Lucene versteht, neben einer einfachen Reihe von Termen, auch diverse im Information Retrieval übliche syntaktische Konstruktionen, die sich untereinander frei kombinieren lassen (vgl. [22], Seite 74):

Boolesche Logik: Die bereits in Abschnitt 2.1.1 ab Seite 10 beschriebene Systematik der Verknüpfung von einzelnen Anfrageteilen durch die logischen Operatoren AND, OR und NOT unterstützt Lucene auch in der Kurzform mit Plus- und Minuszeichen. Entsprechend bewirkt `Deutschland AND Berlin` dasselbe wie `+Deutschland +Berlin`.

Gewichtung: Bei der Anwendung von Lucene ist es möglich, während der Indexierung bestimmte Felder, oder während des Suchprozesses bestimmte Terme oder Phrasen anders zu gewichten, als das einfache Standardmaß. Um einen Term einer Anfrage in seinem Gewicht für die Anfrage zu verändern, wird ein Circonflex und das zu verwendende Gewicht dem Term nachgestellt wie im Beispiel `Frankreich^2,5`. In diesem Beispiel wird der Begriff Frankreich zweieinhalb Mal so stark gewichtet wie alle anderen Terme mit normaler Gewichtung.

Angabe von Feldnamen: Durch das Voranstellen eines Feldnamens ist es in der Anfrage möglich, einen Term oder eine Phrase auf ein anderes Feld als das Standardfeld zu beziehen. Entsprechend resultiert die Anfrage `title:Berlin` in einem Suchprozess, der nur die Titelfelder der Dokumente berücksichtigt.

Gruppierung durch Klammern: Durch das Setzen von Klammern um mehrere Terme wird ein vorangestellter wie z.B. eine Feldanweisung oder folgender Befehl, z.B. eine Gewichtung, auf alle Elemente in der Klammer bezogen. Die Aussage `abstract:(Europa Wirtschaft)^3` wird also aufgelöst zu `abstract:Europa^3 abstract:Wirtschaft^3`.

Exakte Phrasen: Durch das Einschließen mehrerer Terme in Anführungszeichen wird

nach der exakten Phrase gesucht. Die Anfrage „`China und Europa`“ findet ausschließlich Dokumente, in denen diese Wortfolge exakt abgebildet ist.

Distanzangabe: Durch die Kombination von Anführungszeichen und einer nachfolgenden Tilde sowie einem Integer-Wert wird festgelegt, wie viele Wörter die gegebenen Terme in einem Text auseinander stehen dürfen. „`China Europa`“`~1` hätte entsprechend einen ähnlichen Effekt wie das vorherige Beispiel für exakte Phrasen, hier können die Terme aber auch in umgekehrter Reihenfolge vorkommen.

Diese Funktionen werden im Rahmen der Entwicklung der Evaluierungsläufe angewendet, drei Beispiele für konstruierte Anfragen werden in den Tabellen 9.3, 9.4 und 9.5 ab Seite 63 gezeigt.

6.2. Digester

Jakarta Commons Digester ist eine Java-basierte Klassenbibliothek, die das Parsen von XML-Dateien realisieren kann. Digester wandelt dabei die Elemente der XML-Struktur in Java-Objekte um, die wiederum mit Lucene in einen Index geschrieben werden können. Digester lässt sich zur Erkennung von XML-Elementen über diverse Formen von Regeln steuern, die die XML-Architektur abbilden. Beispielhaft seien die folgenden Zeilen der Klasse *DataIndexer* aufgeführt, die drei verschiedene Elemente des XML-Datensatzes in Java Objekte wandeln und an entsprechende Methoden übergeben:

```
digester.addCallMethod("Collection/Publication/Subject/Text", "setControlledVocabulary", 0);
digester.addCallMethod("Collection/Publication/Language/Text","setLanguage", 0);
digester.addCallMethod("Collection/Publication/Description/Text","setAbstract", 0);
```

Eine detaillierte Dokumentation zur Verknüpfung von Digester und den Indexierungskomponenten von Lucene, die im Rahmen der Entwicklungsarbeit für das Indexierungsmodul des vorgestellten IR-Systems adaptiert wurde, findet sich unter [21].

7. Indexierung

Der Datenbestand des FIV ist ein, wie in Abschnitt 4.1.4 beschrieben, äußerst umfassend mit Deskriptoren ausgestattetes Verzeichnis von Dokumenten. Im Rahmen der Indexierung der Datendateien ist es entsprechend ein zentrales Ziel, möglichst viele Metainformationen typenspezifisch im Index zu verzeichnen, um sie später durch das Entry Vocabulary Modul einsetzen zu können.

Das XML-Format des FIV ist im beispielhaften Vergleich zum GIRT-Datensatz äußerst komplex (vgl. auch Abschnitt 4.1.2). Im Rahmen der Entwicklungsarbeit der Indexierungsklassen ist es also notwendig, die für die Retrievalleistung wichtigsten XML-Felder zu identifizieren, gegebenenfalls ähnliche Felder zu kombinieren und so entsprechend jedes verzeichnete Dokument im Index abzulegen.

7.1. Indexierung der Datendateien

Die Indexierung der insgesamt 600 XML-Datendateien des Fachinformationsverbundes geschieht mit Hilfe des XML-Parser Jakarta Commons Digester. Es wird auf die folgenden Felder indexiert:

Indexfeld	XML-Element	Beschreibung
file	—	Dateiname der XML-Datei
id	Collection/Publication/Identifier/Text	Eindeutige Dokumentennummer
title	Collection/Publication/Text	Dokumententitel
abstract	Collection/Publication/Description/Text	Zusammenfassung
subject	Collection/Publication/Subject/Text	Thematische Deskriptoren
language	Collection/Publication/Language/Text	Sprache des Dokuments
classification	Collection/Publication/Classification/Text	Klassifikationsangaben
geo	Collection/Publication/GeographicCoverage/Text	Geopolitische Deskriptoren
temp	Collection/Publication/TemporalCoverage/Text	Temporale Deskriptoren

Tabelle 7.1.: Felder des Datenindex

Die Einträge der Felder des kontrollierten Vokabulars wurden dabei um ihren klassifikatorischen Präfix gekürzt, da zum Zeitpunkt der Indexierung ebenfalls eine Anwendung des kontrollierten Vokabulars auf die freien Felder erwogen wurde, die dann aber im Rahmen der Evaluierung nicht umgesetzt wurde.

7.2. Indexierung der Thesaurusdateien

Auch das Parsing der neun XML-Dateien des Thesaurus wird mit Hilfe von Digester realisiert, es wird auf die folgenden Felder indexiert:

Indexfeld	XML-Element	Beschreibung
subject	Collection/Subject/Text	Thematische Deskriptoren
translations	Collection/Subject/Subject/Text	Nach Typerkennung: Übersetzung
group	Collection/Subject/Subject/Text	Nach Typerkennung: Klassifikationsangaben
subGroup	Collection/Subject/Subject/Text	Nach Typerkennung: Zus. Klassifikationsangaben
connectedTerms	Collection/Subject/Subject/Text	Nach Typerkennung: Zusätzliche Deskriptoren

Tabelle 7.2.: Felder des Thesaurusindex

Der Hinweis auf die Typenerkennung bezieht sich auf eine vereinfachte Lösung zur Indexierung von XML-Elementen, die verschiedene Arten von Metainformationen enthielten. So wurden in der Elementebene Collection/Subject/Subject/Text beispielsweise sowohl Übersetzungen der Deskriptoren, als auch Klassifikationsangaben gespeichert. Die XML-Elemente unterschieden sich dabei nur in ihren Attributen. Das Parsing nach Attributen ist durch Jakarta Commons Digester allerdings nur umständlich zu realisieren. Entsprechend wurde in der Klasse *ThesaurusIndexer* eine Methode entwickelt, die die einzelnen Typen von Metainformationen anhand der codierten Präfixe der Deskriptoren unterscheiden und entsprechend auf verschiedene Felder indexieren kann.

7.3. Zusätzliche Informationen zum Indexierungsprozess

Für die Indexierung wird der für westliche Sprachen geeignete Lucene StandardAnalyzer als Stemmer mit Stoppworten verwendet. Die Stoppwortliste wurde aus den besonders hochfrequent auftretenden Termen des Index mit Hilfe der Lucene Index Toolbox Luke in Version 0.6 entwickelt und zusätzlich ansatzweise auf die Evaluierungsanfragen angepasst.

Sowohl für die Indexierung der Datendateien als auch des Thesaurus wird bei mehreren Einträgen in der gleichen XML-Elementebene eines Dokuments das entsprechende Feld des Lucene Index um alle weiteren Einträge erweitert. Hierbei wird ein Trennzeichen (#) zwischen den Phrasen eingefügt, das während der statistischen Analyse entfernt wird und keinen weiteren Einfluss auf die Indexierung hat.

8. Architektur des Suchprozesses

Um das Entry Vocabulary Modul zu evaluieren, wurde ein Retrievalsystem entworfen, dass bereits über übliche Funktionen wie ein Blind Relevance Feedback Modul und eine Übersetzungsfunktion verfügt. Die Module werden in der Evaluierung auch einzeln in Ihrer Leistung verglichen (vgl. Abschnitt 10.1.4 ab Seite 82). In diesem Rahmen sei darauf hingewiesen, dass die beiden Module der Übersetzung und des Blind Relevance Feedback aufgrund der ausschließlichen Ausrichtung auf die Indexfelder der Titel und der Zusammenfassungen und durch eine vergleichsweise geringe Gewichtung so limitiert worden sind, dass sich die Leistungsfähigkeit der Module in den Protokollen der Evaluiernugsläufe überprüfen lässt, sie grundsätzlich aber nur geringere Auswirkungen erzielen können - der Fokus soll letztendlich auf dem EVM liegen. Im Folgenden werden die einzelnen Schritte des Suchprozesseses entsprechend der SwpEvm-Evaluierungsläufe beschrieben.

Der Suchprozess verläuft im entwickelten System in einzelnen Etappen. Nacheinander wird zunächst die Suchanfrage auf sinntragende Elemente reduziert (Discriminator Modul, vgl. Abschnitt 8.2) und dann die reduzierte Anfrage mit Hilfe des Thesaurus übersetzt (Translator Modul, vgl. Abschnitt 8.3). Die reduzierte Anfrage wird daraufhin entsprechend um das Ergebnis der Übersetzung, des Blind Relevance Feedback (vgl. Abschnitt 8.4) und des Entry Vocabulary Moduls (vgl. Abschnitt 8.5) erweitert. Einen gesamten Überblick des Ablaufs zeigt Abbildung 8.1.

Die eigentliche und abschließende Suche wird zum Ende des gesamten Prozesses mit Hilfe der, durch die einzelnen Schritte augmentierten, Anfrage durchgeführt und das Ergebnis des Suchprozess zu Evaluierungszwecken protokolliert. Es werden insgesamt pro Evaluierungsanfrage die 200 am höchsten bewerteten Dokumente beurteilt.

8.1. Vorbereitende Maßnahmen

Das vorgestellte Retrievalsystem entfernt in einem vorgelagerten Bearbeitungsschritt alle Interpunktionszeichen, um eventuelle syntaktische Funktionen der Anfragenverarbeitung von Lucene nicht zusätzlich anzusprechen. Darüber hinaus werden auf diesem Wege in mehreren Fällen durch Zeichen verbundene Terme voneinander entkoppelt.

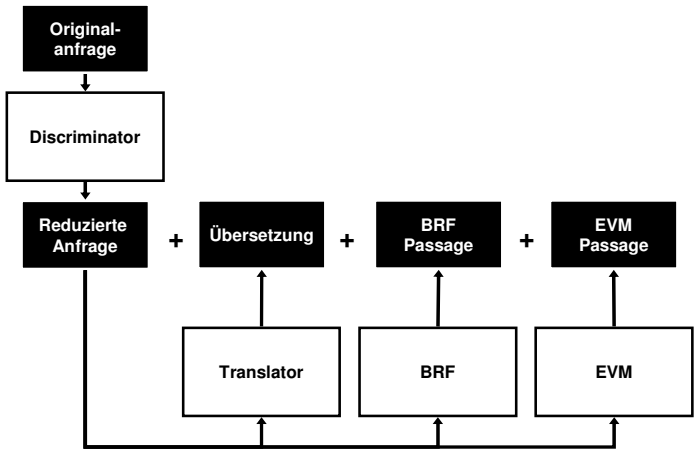

Abbildung 8.1.: Ablauf der Anfragenerweiterung

8.2. Discriminator

Um aus der natürlichsprachlichen Anfrage die sinntragenden Elemente herauszufiltern und so eine sowohl im Umfang reduzierte als auch inhaltlich verdichtete Anfrage zu formulieren, wird im vorliegenden Retrievalsystem das Discriminator Modul eingesetzt.

Die Aufgabe des Discriminator Moduls ist es, Begriffe, die bereits beim Indexieren durch die Stoppwortliste ausgeschlossen wurden und Begriffe, die nicht während der Indexierung ausgeschlossen wurden, und besonders häufig im Index vorkommen, aus der Anfrage zu entfernen.

Hierbei ist das Ziel, dass das System hier nicht eine semantisch zentrale, freie Vokabel aus der Anfrage entfernt, nur weil sie nicht im kontrollierten Vokabular der Metadaten der einzelnen Dokumente vorkommt. Darum sucht das Discriminator Modul nach jedem einzelnen Term der Anfrage im Datenindex auf die freien Suchfelder *abstract* und *title*. Ergibt sich mindestens ein Treffer in einem der beiden Felder, verbleibt der Term in der weiter zu verwendenden Anfrage. Ergibt sich bei der Suche in den Daten bei einem Term eine Trefferliste mit mehr als 8000 Einträgen (Zahl auf Basis von Tests mit SWP-Evaluierungsanfragen festgestellt, Schwellenwert knapp über dem mit rund 7500 Nennungen am häufigsten verzeichneten Begriff „China“ von allen Begriffen der Evaluierungsanfragen), wird dieser Term im weiteren Suchprozess ignoriert. Diese Reduktion der Anfrage ist besonders für den Übersetzungsprozess im Translator Modul notwendig, da ohne eine solche Verkürzung oftmals versucht würde, sinnfreie Teile der Anfrage zu übersetzen und somit die Anfrage weiter um semantisch irrelevante Passagen zu ergänzen.

8.3. Translator

In [36] beschreibt Petras erfolgreiche Versuche der Übersetzung mit Hilfe eines mehrsprachigen Thesaurus mit einem zusätzlichen Hinweis auf [38]. Entsprechend wurde versucht, auch in diesem Rahmen eine Übersetzung von Anfragetermen durch den Thesaurus zu realisieren. Da die Übersetzungsleistung in einem Retrievalsystem für eine spezifische, inhaltliche Domäne zu großen Teilen von der thematischen Eignung des eingesetzten Wörterbuchs abhängt, war die Nutzung der Übersetzung des vorliegenden Thesaurus mit seinen Übersetzungen naheliegend und vielversprechend.

Dem Translator Modul wird die vom Discriminator Modul reduzierte Suchanfrage übergeben. Das Übersetzungsmodul durchsuch daraufhin das Feld *subject* des Thesaurus und gibt bei einem Score von 1,0, also einer exakten Übereinstimmung, die Inhalte des Feldes *translations* zurück. Da das parallel entwickelte Modul zur Erkennung von Phrasen nicht rechtzeitig fertiggestellt werden konnte, muss sich das Translator Modul entsprechend auf termweise Übersetzungen beschränken, auch wenn bewusst ist, dass die Übersetzung von Phrasen mitunter bessere Ergebnisse erzielen kann.

Trotzdem hat eine solche Methode zur Übersetzung grundlegendes Potential: Beispielsweise lassen sich eintermige Ländernamen und Themenangaben, die zentralen Charakter für den Sinn einer Anfrage beinhalten, auf diese Art und Weise zuverlässig in Englisch, Französisch und Spanisch in der für die Datenbasis passenden Fachterminologie übersetzen.

Darüber hinaus sei darauf hingewiesen, dass der Einsatz des im Abschnitt 8.5 beschriebene EVM in einem IR-System für die Datenbank des FIV einen Großteil der mehrsprachigen Retrievalleistung realisiert, da das kontrollierte Vokabular über die Dokumente aller Sprachen in einer einheitlichen Sprache verfasst ist und sich somit mit Hilfe der Deskriptoren ein grundlegendes, mehrsprachiges Retrieval realisieren lässt.

Die Ergebnisse der Übersetzung wird der Anfrage in Form des durch Lucene unterstützten Wortabstands-Syntax ergänzt (beispielsweise „Vereinigte Staaten“~1, vgl. Abschnitt 6.1 auf Seite 42), was grundsätzlich dem Syntax der exakten Phrase entspricht. Die Gewichtung bleibt einfach.

8.4. Blind Relevance Feedback

Das Blind Relevance Feedback Modul (BRF) wurde nach geringfügiger Anpassung auf den erzeugten Index aus dem System der Universität Hildesheim übernommen. Es wurde bereits in mehreren Systemen erprobt, vgl. beispielhaft [24].

Während das im folgenden Kapitel vorgestellte EVM das lokale Feedback im Bezug auf Metadatenfelder realisieren soll, wird das BRF zusätzlich eingesetzt, um auch die Freitext-Felder *title* und *abstract* für Relevance Feedback zu nutzen. Auf diese Weise werden alle zur Verfügung stehenden Felder durch die beiden verschiedenen Varianten des Feedbacks genutzt.

Dem Blind Relevance Modul wird ebenfalls die durch den Discriminator reduzierte Anfrage übergeben. Im Rahmen der Evaluierung werden bei Einsatz des BRF pro Anfrage die 30 am besten bewerteten Dokumente untersucht und fünf Terme zur Anfragenergänzung zurückgegeben und auf die Felder *title* und *abstract* gerichtet. Die beiden Werte haben sich im Rahmen einer vorbereitenden Erprobung als vergleichsweise geeignet herausgestellt. Es wird die Berechnungsmethode des Robertson Selection Value verwendet (vgl. Abschnitt 2.1.3).

8.5. Entry Vocabulary Modul

Im Entry Vocabulary Modul werden, entsprechend der detaillierten Beschreibung in Abschnitt 5 ab Seite 36, in diversen Durchgängen mit Hilfe der durch den Discriminator reduzierten Anfrage Metadaten aus den Feldern des kontrollierten Vokabulars extrahiert und der Anfrage angehängt.

Es wurde also nach Abschnitt 5.4 ein dynamisches EVM mit anfragenorientierter, mehrdimensionaler und kaskadenförmiger Extraktion von Metadaten umgesetzt. Die Umsetzung der anfragenorientierten Extraktion von Metadaten geht auf das Konzept der Anfragenerweiterung zurück - in diesem Modell wird die Anfrage nicht in das kontrollierte Vokabular übersetzt, sondern um diverse Deskriptoren ergänzt. Der mehrdimensionale Ansatz empfiehlt sich, da der Datenbestand über mehrere Bezugssysteme von Deskriptoren erschlossen ist. Die kaskadenförmige Wiederverwendung von extrahierten Termen wird deshalb eingesetzt, da nur ein vergleichsweise kleiner Anteil von Dokumenten über Zusammenfassungen verfügt und somit direkt über die Anfrage auf die Felder des freien Vokabulars auswertbar ist. Entsprechend wird die umfassende Erschließung durch Metadaten in einer zweiten Generation von Anfragen durch die bereits extrahierten Deskriptoren genutzt und so der gesamte Datenbestand per Deskriptorensuche angesprochen und ausgewertet.

Der Prozess dieses EVM ist in Abbildung 8.2 abgebildet: Zunächst wird die reduzierte Anfrage an die Klasse *EntryVocabularyModule* übergeben, dann, in den mehreren Schritten der Kaskade in der Klasse *ControlledVocabularyAdder* mit Hilfe der Klasse *Searcher* auf diverse Indexfelder angewendet und die Ergebnisse statistisch ausgewertet. Die am höchsten bewerteten Terme und Phrasen werden an die Klasse

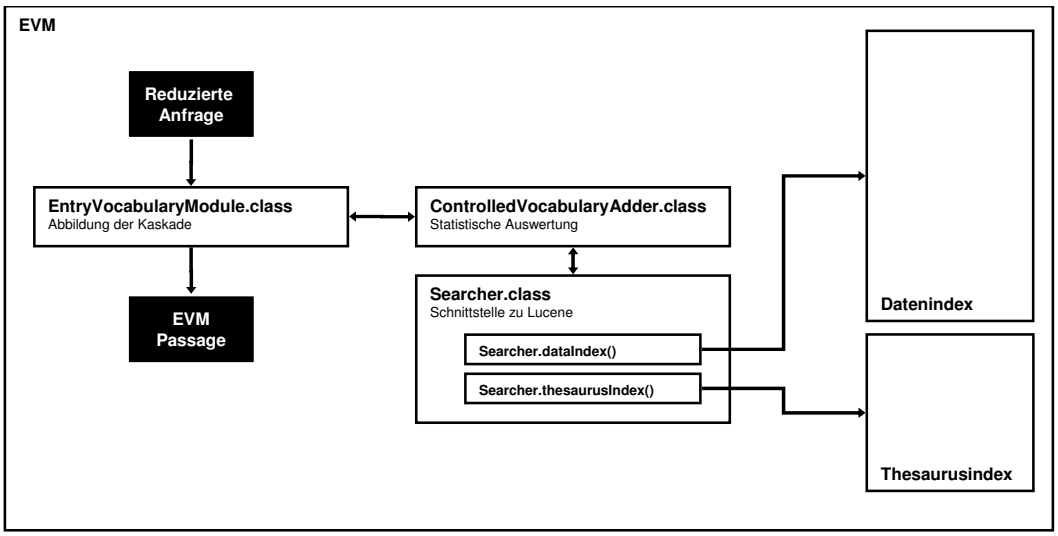

Abbildung 8.2.: Prozessablauf des EVM

EntryVocabularyModule übergeben, wo alle Einzelergebnisse der Kaskadenelemente zusammengefasst und zur Ergänzung der ursprünglichen Anfrage zurückgegeben werden.

Im Falle der vorliegenden Datenbasis sind die drei für die Extraktion relevanten Felder der Datenbasis *geo*, *subject* und *classification*. Darüber hinaus werden die Felder *subject* und *connectedterms* des Thesaurus für eine Extraktion berücksichtigt. Die beiden freien Felder der Datenbasis, auf die zunächst die reduzierte Anfrage gerichtet wird sind *title* und *abstract*.

Das vorgestellte Modell wendet eine Anfragenkaskade an, die sowohl die reduzierte Anfrage, als auch aus der reduzierten Anfrage gewonnene Deskriptoren zur Gewinnung von weiteren Metadaten verwendet. Einen genauen Überblick über den Verlauf der Kaskade im evaluierten System gibt Abbildung 8.3 und die entsprechende Tabelle 8.1. In der Tabelle werden detailliert die verwendete Anfrage, die Richtung der Anfrage auf das gegebene Feld und das ausgewertete Feld der gefundenen Dokumente sowie der zurückgegebene String aus potentiell nützlichen Deskriptoren genannt. Die Anfragenkaskade ist in der Klasse *EntryVocabularyModule* programmiert.

Im Rahmen der statistischen Auswertung wird das Suchergebnis jedes Kaskadenelements in *ControlledVocabularyAdder* untersucht. Der Umfang der Untersuchung lässt sich durch den Faktor *consideredDocs* steuern: Hier wird angegeben, wie viele der am höchsten bewerteten Dokumente in die statistische Auswertung eingehen. Die Evaluierung berücksichtigt hier die Parameter 30 und 100 Dokumente. Über die in *consideredDocs* genannte Zahl von Dokumenten werden alle Terme und Phra-

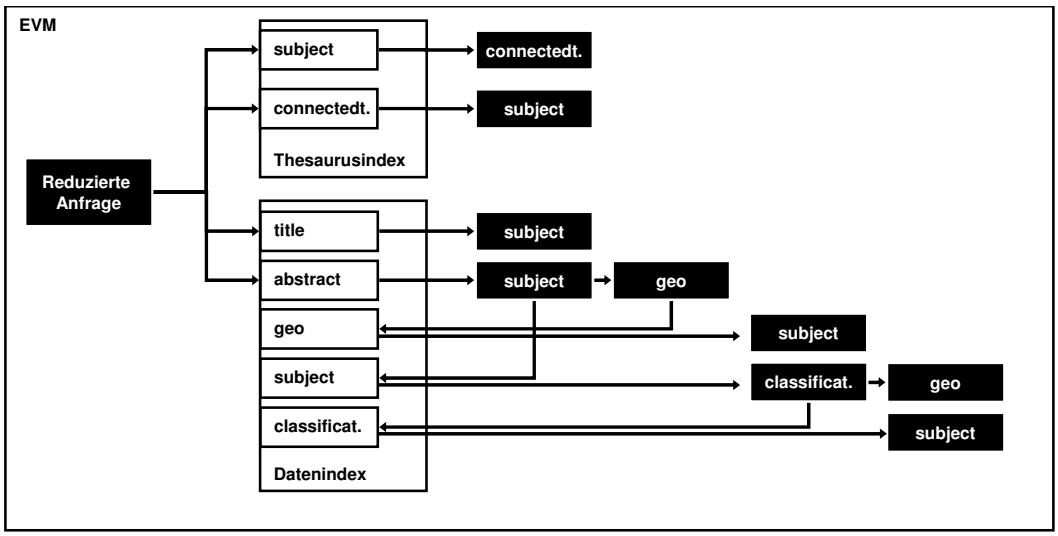

Abbildung 8.3.: Verlauf der EVM-Kaskade

	Anfrage	Anfragefeld	Rückgabefeld	Ergebnisstring
An den Thesaurus gerichtete Anfragen:				
1.	stoppedQuery	subject	connectedterms	evmCtFromSub
2.	stoppedQuery	connectedterms	subject	evmSubFromCt
An den Datenindex gerichtete Anfragen:				
3.	stoppedQuery	title	subject	evmSubjectsFromTitleQuery
4.	stoppedQuery	abstract	subject	evmSubjectsQuery
5.	stoppedQuery	abstract	geo	evmGeoQuery
6.	evmGeoQuery	geo	subject	evmSubjectsFromGeoQuery
7.	evmSubjectsQuery	subject	classification	evmClassFromSubjectsQ.
8.	evmSubjectsQuery	subject	geo	evmGeoFromSubjectsQuery
9.	evmClassFromSubjectsQ.	classification	subject	evmSubjectsFromClassQ.

Tabelle 8.1.: Detaillierter Verlauf der Kaskade

sen des untersuchten Felds mit dem Score ihres Ursprungsdokuments verknüpft. Bei Mehrfachnennungen werden die Werte addiert. Über 30 Dokumente werden so z.B. im Feld *subject* bei durchschnittlich 12 Deskriptoren pro Dokument 360 Deskriptoren untersucht und ausgewertet. Abschließend werden die addierten Scores der einzelnen Deskriptoren normalisiert. Bevor die Deskriptoren zurückgegeben werden, nehmen die Parameter *cutOffScore* und *numberOfReturned* Einfluß auf die Anzahl der zurückgegebenen Dokumente. *cutOffScore* wird in der Evaluierung mit 0,3 und 0,6 getestet, dies bedeutet, dass hier nur Deskriptoren verwendet werden, die mindestens einen Score von 0,3 oder 0,6 erreichen. *numberOfReturned* limitiert die Anzahl der maximal zurückgegebenen Deskriptoren, falls sehr viele Deskriptoren ein höheres Ergebnis erzielt haben als in *cutOffScore* gefordert. In der Evaluierung wurden durchweg maximal sechs Terme oder Phrasen ergänzt. Hintergründe der Parameterwahl werden in

Abschnitt 9.3.3 ab Seite 66 beschrieben.

Im Rahmen von *ControlledVocabularyAdder* wird ebenfalls der Syntax der zurückgegebenen Terme und Phrasen festgelegt. In zwei Evaluierungsläufen wird dabei zusätzlich mit dem Deskriptor der normalisierte Score aus der statistischen Auswertung als Gewicht verknüpft.

8.6. Abschließende Maßnahmen

Das Retrievalsystem gibt nach jedem Evaluierungslauf zwei Ergebnisdateien aus. Einerseits wird in einem standardisierten Format die Auflistung der gefundenen Dokumente über alle eingelesenen und bearbeiteten Anfragen in der Results-Datei dokumentiert, andererseits zeichnet die Output-Datei die während der Verarbeitung der Anfrage vorgenommenen Veränderungen und Ergänzungen jedes einzelnen Moduls und abschließend die tatsächlich eingesetzte Anfrage auf.

Die Results-Datei folgt dem im Rahmen von Evaluierungsprojekten etablierten TREC-Standard, eine entsprechende Methode wurde aus einer Quellcode-Datei des Systems der Universität Hildesheim übernommen und an die Gegenbenheiten angepasst. Beispielhaft wird eine Zeile aus der Results-Datei des SwpEvm1- Evaluierungslaufs gezeigt:

```
1 Q0 swp-fiv-d444794 24 0.41883114 SwpEvm1_25q swp_fiv_448.xml
```

Die Angaben je Zeile sind in Tabelle 8.2 aufgeschlüsselt.

Passage	Beispiel	Beschreibung
1	1	Nummer der Suchanfrage (1-25)
2	Q0	Nummer der Iteration (immer Q0, da kein iteratives System)
3	swp-fiv-d444794	Dokumentennummer
4	24	Rang des Dokuments in der Auflistung (1-200)
5	0.41883114	Lucene Score (0-1)
6	SwpEvm1_25q	Bezeichnung des Evaluierungslaufs
7	swp_fiv_448.xml	Dateiname der Datei, in der das Dokument verzeichnet war

Tabelle 8.2.: Felder der Einträge in den Results-Dateien

Die Output-Datei bildet die einzelnen Teile und die tatsächlich im Suchprozess verwendete Anfrage in einem freien Format ab. Die Datei wird für jeden einzelnen Evaluierungslauf generiert und ist zur Analyse der Erweiterung der Suchanfrage hilfreich. Im Folgenden wird beispielhaft ein Ausschnitt der Output-Datei des SwpEvm2-Laufs gezeigt.

```
Query # : 3
ENTERED QUERY >> Welche Faktoren bestimmen die Beziehungen zwischen China...
STOPPED QUERY >> Faktoren Beziehungen China EU EU
TRANSLT QUERY >> "‘Beziehungen"~1 "‘Relationship"~1 "‘relations"~1 "‘relacione...
BLINDRF QUERY >> mitgliedsländer china akteur regionale prozesse
EVMODUL QUERY >> abstract:(Europäische Strukturfonds)^1.0 abstract:(Bestim...
CONSTRC QUERY >> title:(Faktoren Beziehungen China EU EU )^10 abstract:(Fa...
```

Teil IV.

Evaluierung

9. Vorbereitungen der Evaluierung

Nachdem in den vorhergehenden Kapiteln die Grundlagen gelegt und das entwickelte System beschrieben wurde, wird in diesem Kapitel zunächst der Evaluierungsprozess ausgehend von der organisatorische Realisierung (vgl. Abschnitt 9.1) beschrieben. Darauf folgen eine Analyse der Evaluierungsanfragen in Abschnitt 9.2 und die Entwicklung der Evaluierungsdurchgänge in 9.3 ab Seite 60. Abschnitt 9.4 auf Seite 69 beschließt das Kapitel mit der Beschreibung der Evaluierung des mehrsprachigen Retrievalergebnisses.

9.1. Realisierung der Evaluierung

Da es für den Datenbestand des Fachinformationsverbundes keine umfassenden Relevanzbewertungen zu den vorliegenden Fragestellungen gibt, basiert die Evaluierung der durch das System erzielten Ergebnisse auf einer auf die gefundenen Dateien begrenzte Relevanzbewertung und deren Auswertung.

Die Relevanzbewertung wurde von Fachreferenten des Fachinformationsbereichs der SWP und von Studenten des Internationalen Informationsmanagements mit dem Schwerpunkt Informationswissenschaften und des Nebenfachs Politik der Universität Hildesheim durchgeführt. Hierbei übernahm die SWP die Suchanfragen 6-25, die Anfragen 1-5 wurden durch die Universität Hildesheim getragen.

9.2. Evaluierungsanfragen

Im Vergleich zu den in diversen IR-Foren wie TREC oder CLEF etablierten Anfrageformaten mit mehreren Feldern, wurde durch die SWP eine einfache Sammlung von Anfragesätzen bereitgestellt. Im Anhang ab Seite 108 werden die Evaluierungsanfragen abgebildet. Eines der zentralen Argumente für in mehreren Zeilen verfasste, detaillierte Beschreibung der gesuchten Dokumente, wie beispielsweise in den Evaluierungsanfragen der Datenbasis GIRT, ist die bessere Grundlage zur Relevanzbewertung. Der mit

der Relevanzbewertung betraute Nutzer ist anhand einer umfassenden Beschreibung der gesuchten Dokumente besser in der Lage zu unterscheiden, ob die aufgelisteten Dokumente zu der Gruppe der Relevanten gehört.

Die Evaluierungsanfragen wurden im Rahmen der Evaluierung ebenfalls auf Anhaltspunkte hin analysiert, die anfragenorientierte Untersuchung findet sich im Anhang ab Seite 110, die Ergebnisse der Untersuchung werden im Folgenden kurz zusammengefasst.

9.2.1. Unterschiedliche Typen der Informationsangaben

Die Informationsangaben in den Anfragen lassen sich in drei Gruppen aufteilen:

Der erste Typ, der thematische Bezug, kommt in jeder Evaluierungsanfrage vor. Beispielhaft für solche thematischen Bezüge können die konkreten Begriffe „Minderheitenpolitik“ (Anfrage 12), „Wettbewerbspolitik“ (Anfrage 19) und „Rechtsextremismus“ (Anfrage 20) genannt werden. Es gibt darüber hinaus auch recht unkonkrete und allgemeine thematische Bezüge wie „Beziehungen“ (Anfragen 1, 3, 5, 6, 7, 11) oder „Haltung“ (Anfrage 15).

Der zweite Typ, der geopolitische Bezug, kommt in 88% der Evaluierungsanfragen (bis auf Anfragen 5, 8, 24) vor. Beispielhaft für solche geopolitschen Bezüge können die konkreten Ländernamen „USA“, „Libyen“ und „China“ genannt werden. Außerdem kommen aber auch unpräzisere geopolitische Angaben wie Regionenbezeichnungen in einigen Fragen vor, so zum Beispiel „asiatisch-pazifischen Raum“ (Anfrage 6), „Westlicher Balkan“ (Anfrage 11) oder „mediterranen Raums“ (Anfrage 23).

Der dritte Typ, der temporale Bezug, kommt nur in 20% der Evaluierungsanfragen (1, 2, 18, 22 und 23) vor. Beispielhaft für temporale Bezüge können die Angaben „in den letzten zehn Jahren“ (Anfragen 1 und 22) und „aktuelle“ (Anfrage 2) genannt werden. Kompliziertere Bezüge sind in den Anfragen 18 - zum Entschlüsseln der Angabe „nach Abschluss der Uruguay-Runde“ ist das Wissen über den Zeitpunkt einer Veranstaltung notwendig - und 23 - der Bezug auf die „Zukunft“ lässt sich üblicherweise nicht in Form von kontrolliertem Vokabular ausdrücken - enthalten. Dieser dritte Typ wird im weiteren Verlauf der Entwicklung nicht beachtet, aber als Ansatz für eine Weiterentwicklung im Ausblick im Abschnitt 11.2.4 auf Seite 100 betrachtet.

9.2.2. Unterschiedliche Konkretisierungsgrade

Es lässt sich feststellen, dass alle Fragen unterschiedlich konkret formuliert sind. Während die Fragen 2, 9 und 10 Beispiele für konkrete Verknüpfungen von einem Themenkomplex („Wiederaufbau“, „Erdölpolitik“, „Atomkonflikt“) und einem geopolitischen

Bezug („Afghanistan“, „Libyen“, „China“) darstellen, können die Fragen 11, 18 und 23 als Beispiele für vergleichsweise offen formulierte Fragen genannt werden. Im Folgenden soll versucht werden, die Anfragen nach Konkretisierungsgrad zu ordnen, um später in Abschnitt 10.1.2 ab Seite 73 im Rahmen einer anfrageorientierten Evaluierungsauswertung Schlüsse ziehen zu können.

Als Anhaltspunkt für die Bewertung, ob ein Anfragefragment konkret und trennscharf oder aber vage und uneindeutig definiert ist, gilt bei Termen und Phrasen mit geopolitischem Bezug eine Skala, die definiert sei von der Angabe von konkreten Ländernamen („Nigeria“) über Namen von Staatenbünden („EU“) über Bezeichnungen für Regionen („Mittlerer Osten“) bis hin zu der Anwendung von vagen und offeneren Bezeichnung von größeren, regionalen Räumen („mediterraner Raum“). Dieser Maßstab wurde gewählt, da er im Kontext der internationalen Beziehungen die Skala von der eigentlichen Ebene der Internationalität, der Ebene der einzelnen Staaten, über mehrere Ebenen von supranationalen Organisationen bis hin zu regionalen Bezugsystemen und transkontinentalen Verbünden sowohl eine entsprechende Entwicklung der Anzahl der beteiligten Staaten von einem Staat bis hin zu Gruppen von mehreren Dutzend Staaten als auch eine eine Entwicklung der eindeutigen Definition vom klar abgegrenzten Staat über einen klar definierten Staatenbund bis hin zum unklar definierten Regionalverbund abbildet.

Bei Begriffen, die einen thematisch-inhaltlichen Bezug haben, wurde eine Bewertung gewählt, die grundsätzlich der Hierarchie einer Klassifikation von politischen Themen abbildet. Während gänzlich offene Begriffe wie „Probleme“ und „Herausforderungen“ hierbei das eine, vage Ende der Skala beschreiben, begrenzen die konkreten Schlagworte wie „Entwaffnung“ oder „Massenvernichtungswaffen“ das andere Ende der Skala. Zwischen den beiden Extremen der Skala finden sich oberbegriffsartige Themenangaben wie beispielsweise „Wettbewerbspolitik“ oder „Minderheitenpolitik“.

Abbildung 9.1 verdeutlicht die Aufteilung der Anfragen über verschiedene Grade der Konkretisierung über den geopolitischen und den thematischen Bezug nach diesem Schema. Bei mehreren verschiedenen Konkretisierungsgraden in einer Anfrage wurde die Frage der jeweils weniger konkreten Kategorie zugeordnet.

Während die konkreten, geschlossenen Fragen im Bezug auf das Retrievalergebnis wegen eindeutiger Schlagworte mehr Erfolg versprechen, sind die offenen und freien Anfragen besonders schwierig in nutzbare Retrievalergebnisse umzusetzen: In Frage 2 lassen sich ohne aufwändige Ergänzungsmethoden zwei zentrale Schlagworte (Wiederaufbau, Afghanistan) herausfiltern, die bei einer Suche erfolgreich eingesetzt werden können. Frage 3 hingegen gibt bis auf die gänzlich offenen Begriffe „Faktoren“ und „Beziehungen“ keine direkten Anhaltspunkte für die Inhalte, die ein relevantes Dokument

enthalten sollte. Andererseits ist das Ergebnis einer offenen Anfrage auch ein größerer Raum an potentiell relevanten Dokumenten. Es ist also interessant zu untersuchen, ob sich die Retrievalleistung zum einen, oder anderen Ende der Skala hin auch im Vergleich zwischen den einzelnen Evaluierungsläufen (vgl. Abschnitt 9.3) verändert.

Es wird zu untersuchen sein, inwiefern die konkreten und die vagen Anfragen unterschiedliche Ergebnisse im vorliegenden Retrievalsystem erzielen. Bei der Auswertung kann Abbildung 9.1 hilfreich sein, um zu bewerten, in welchen Feldern verschiedene Evaluierungsläufe erfolgreicher waren als in anderen.

9.2.3. Mehrfache Kontexte in einer Anfrage

Einige Anfragen sind so formuliert, dass sie mehrfache Anfragekontexte enthalten. Beispielhaft kann Anfrage 3 genannt werden, die sowohl die Beziehungen zwischen der EU und China als auch allen anderen EU-Staaten und China zum Thema hat. Diese mehrfachen Anfragekontexte sind in sich äußerst heterogen und erfordern grundsätzlich die Formulierung von mehreren Suchanfragen, da sich die relevanten Dokumenten für die einzelnen Anfragekontexte ebenfalls unterscheiden. Insgesamt kommen in den Anfragen 3, 6, 7, 11, 12, 13, 14, 20, 21 und 24 mehrfache und deutlich zu unterscheidende Anfragekontexte vor.

Diese Anfragen unterscheiden sich allerdings im Grad der Abgrenzung der einzelnen Kontexte. Während die Anfragen 11 und 20 beispielsweise gravierend unterschiedliche Kontexte enthalten, sind die Anfragen 6 und 25 vergleichsweise homogen im Kontext. Da sich keine eindeutige Skala ableiten lässt, wird diese Unterscheidung in den Anfragen nicht weiter verfolgt, es handelt sich aber trotzdem um einen bemerkenswerten Aspekt der Gruppe von Evaluierungsanfragen.

	Keine geopolit. Angaben	Nennung von überregionalen Räumen	Nennung von Regionen	Nennung von Staatenbünden	Nennung von Staaten
Präzise Themenangabe	8, 21, 24	7		12	2, 9, 10, 20
Oberbegriffsartige Themenangabe	5, 19, 18	6	4, 13, 25	16, 17, 22	
Vage Themenangabe		11, 23	14	3	1, 15

Abbildung 9.1.: Sortierung der Anfragen nach Systematik aus Abschnitt 9.2.2

9.3. Entwicklung der Evaluierungsdurchgänge

Insgesamt wurden zur Evaluierung des Information Retrieval Systems und des Entry Vocabulary Moduls fünf Evaluierungsdurchgänge entwickelt. Zwei der Läufe, SwpBase1 und SwpBase2 wurden von der SWP vorgegeben und dienen zur Ermittlung einer Referenzlinie, an denen sich die drei weiteren Läufe mit zusätzlichen Modulen wie EVM, Blind Relevance Feedback und Übersetzung messen lassen.

Im Zusammenspiel mit den vorliegenden, natürlichsprachlichen Evaluierungsfragen ergibt sich dadurch eine gute Vergleichsbasis, um zwei für die Unterscheidung der Leistungsfähigkeit des Systems zentrale Situationen zu simulieren: Die Anwendung eines simplen Retrievalsystems und die Anwendung eines weiterentwickelten Retrievalsystems durch einen unerfahrenen Nutzer mit natürlichsprachlichen Anfragen. Im Folgenden werden die fünf Läufe detailliert vorgestellt.

9.3.1. SwpBase-Evaluierungsdurchgänge

Die durch die SWP konzipierten Basis-Evaluierungsdurchgänge sollen dazu verwendet werden, die Veränderungen der Retrievalleistung des Systems während der weiterentwickelten SwpEvm-Evaluierungsläufe zu messen. Die im Folgenden beschriebenen Evaluierungsläufe verfügen also zu Vergleichszwecken über keine besonderen Verbesserungen oder Veränderungen der natürlichsprachlichen Anfrage, bis auf die reguläre Entfernung von jeglicher Interpunktion und die indirekte Anwendung der Stoppworte des Index. Hierbei ist zu beachten, dass die Stoppworte erst beim Suchprozess und nicht tatsächlich bei der Anfragenformulierung entfernt werden: Da der Datenbestand während der Indexierung von Stoppworten befreit wurde, kommt dieses Verfahren aber zum gleichen Ergebnis. Entsprechend werden in der nachfolgenden Übersicht „effektive“ und nicht „tatsächlich“ Anfragen genannt.

SwpBase1 - Erster Basis-Evaluierungslauf ohne Metadaten

Dieser Evaluierungslauf wurde mit dem Ziel entwickelt, mit den Termen der durch die Stoppworte reduzierte Anfrage die Felder *abstract* und *title* zu durchsuchen. Es werden also ausschließlich die verzeichneten Beschreibungstexte und Titel und keine Metainformationen durchsucht.

Tabelle 9.1 zeigt die Anfragenvorbereitung für den Evaluierungslauf am Beispiel einer Anfrage.

	Wortlaut Anfrage
Originalanfrage 1:	Wie haben sich die Beziehungen zwischen Nigeria und den USA in den letzten zehn Jahren entwickelt?
Effektive Anfrage 1:	title:Beziehungen title:Nigeria title:USA abstract:Beziehungen abstract:Nigeria abstract:USA

Tabelle 9.1.: Originalanfrage vs. effektive Anfrage SwpBase1

SwpBase2 - Zweiter Basis-Evaluierungslauf mit Metadaten

Dieser Evaluierungslauf wurde im Gegensatz zum SwpBase1 mit dem Ziel entwickelt, mit den Termen der durch die Stoppworte reduzierte Anfrage die Felder *abstract*, *title*, *subject* und *classification* zu durchsuchen. Es werden also neben den Beschreibungstexten und Titeln auch noch die thematischen Deskriptoren und die Klassifikationsangaben untersucht.

Tabelle 9.2 zeigt die Anfragenvorbereitung für den Evaluierungslauf für Evaluierungsanfrage 1.

	Wortlaut Anfrage
Originalanfrage 1:	Wie haben sich die Beziehungen zwischen Nigeria und den USA in den letzten zehn Jahren entwickelt?
Effektive Anfrage 1:	title:Beziehungen title:Nigeria title:USA abstract:Beziehungen abstract:Nigeria abstract:USA subject:Beziehungen subject:Nigeria subject:USA classification:Beziehungen classification:Nigeria classification:USA

Tabelle 9.2.: Originalanfrage vs. effektive Anfrage SwpBase2

9.3.2. SwpEvm-Evaluierungsdurchgänge

Die drei weiteren Evaluierungsläufe werden auf dem weiterentwickelten IR-System mit aktiviertem Discriminator, Entry Vocabulary Module, Blind Relevance Feedback und der Thesaurus-Übersetzung durchgeführt. Hierbei wird die Anfrage schrittweise um die Ergebnisse der einzelnen Module verändert. Während der Discriminator die Anfrage reduziert, erweitern die anderen Module die Anfrage um Übersetzungen, potentiell relevante Begriffe des Freitextes (BRF) und des kontrollierten Vokabulars (EVM).

Die drei SwpEvm-Evaluierungsläufe unterscheiden sich voneinander in den in Abschnitt 3 beschriebenen Parametern *consideredDocs* und *cutOffScore* des Entry Vocabulary Moduls. Diese Parameter steuern die statistische Auswertung des extrahierten, kontrollierten Vokabulars und haben damit grundlegenden Einfluss auf das Modul und die zurückgegebenen Ergebnisse. Darüber hinaus wurde zwischen SwpEvm1/2 und SwpEvm3 die Syntax des zurückgegebenen Anfrageanteils des EVM geändert:

Während bei den ersten beiden Läufen dem kontrollierten Vokabular ein aus der statistischen Auswertung errechneten Gewicht zugefügt wurde, gehen alle zurückgegebenen Terme des Moduls in SwpEvm3 mit gleichem, einfachen Gewicht ein. Die Ergebnisse des Discriminator Moduls, die reduzierte Anfrage, geht, um den Einfluss von deutlichen Abschweifungen des extrahierten Vokabulars zu limitieren und um den Charakter der Ergänzung (im Gegensatz zur Reformulierung) der ursprünglichen Anfrage durch kontrolliertes Vokabular zu unterstreichen, bei allen drei Läufen mit zehnfachem Gewicht ein. Eine detaillierte Übersicht über alle Evaluierungsläufe und deren Parameter gibt Tabelle 9.6 auf Seite 67.

In den Tabellen 9.3, 9.4 und 9.5 werden beispielhaft drei tatsächliche, durch das System konstruierte Anfragen den ursprünglichen, normalsprachlichen Anfragen gegenübergestellt.

In Abschnitt 10.1.1 auf Seite 72 werden die Ergebnisse der einzelnen SwpEvm-Läufe untereinander verglichen und bewertet. Um diesen internen Vergleich der SwpEvm-Läufe systematisch anzugehen, werden in den folgenden Unterkapiteln die unterschiedlichen Parameter und die theoretischen Auswirkungen zusammengefasst.

SwpEvm1 vs. SwpEvm2

SwpEvm1 unterscheidet sich gegenüber SwpEvm2 in der Veränderung der Anzahl von 30 (SwpEvm1) auf 100 (SwpEvm2) der berücksichtigten Dokumente in der statistischen Auswertung. In dem direkten Vergleich zwischen SwpEvm1 und SwpEvm2 wird sich zeigen, ob die Erhöhung der Anzahl der berücksichtigten Dokumente auch die Retrievalqualität steigern kann.

Dafür spricht, dass durch eine größere Anzahl von ausgewerteten Dokumenten die Anzahl der tatsächlich relevanten Dokumente (wenn auch immer langsamer) steigt und somit der Anteil an statistisch berücksichtigtem, relevanten kontrollierten Vokabular und damit die Gewichtung eines gut geeigneten Deskriptors steigt, während die verschiedenen irrelevanten Deskriptoren unterschiedlicher sind und weniger oft genannt werden, entsprechend weniger durch Mehrfachnennungen ihren Score verbessern können. Während der Entwicklung hat sich gezeigt, dass erst ab einer gewissen Menge von untersuchten Dokumenten die statistische Auswertung durch Mehrfachnennungen von Deskriptoren in mehreren Dokumenten eine ausreichend hohe Bewertung erreicht. Da ein Schwellenwert von 0,3 eingesetzt wird, kann dies die Voraussetzung sein, dass ein geeigneter Deskriptor überhaupt in der Anfrage eingesetzt wird.

Dagegen spricht, dass die die Precision eines Retrievalsystems üblicherweise von Recallstufe zu Recallstufe weiter abnimmt. Entsprechend wäre zu erwarten, dass die Ergebnisse der statistischen Auswertung von deutlich mehr Dokumenten auch deutlich

	Wortlaut Anfrage
Originalanfrage 1:	Wie haben sich die Beziehungen zwischen Nigeria und den USA in den letzten zehn Jahren entwickelt?
Tatsächliche Anfrage 1, generiert im Lauf SwpEvm1:	title:(Beziehungen Nigeria USA)^10 abstract:(Beziehungen Nigeria USA)^10 „Beziehungen“~1 „Relationship“~1 „relations“~1 „relaciones“~1 „Nigeria“~1 „Nigeria“~1 „Nigéria“~1 „Nigeria“~1 „Nigerianer“~1 „Nigerians“~1 „Nigérians“~1 „nigerianos“~1 abstract:(Länder mit Schuldendienstproblemen)^1.0 abstract:(Englischsprachiges Afrika)^1.0 abstract:(AKP-Länder)^1.0 abstract:(Market Borrowers)^1.0 abstract:(Severely indebted low-income countries)^1.0 abstract:(Internationale Beziehungen)^0.919297 subject:(Vereinigte Staaten)^1.0 subject:(Loyalität)^0.9293553 subject:(Multilateral)^0.9293553 subject:(Disengagement)^0.9293553 subject:(Soziale Kontrolle)^0.5808471 subject:(Länder mit Schuldendienstproblemen)^0.5351242 subject:(Vereinigte Staaten)^1.0 subject:(Außenpolitik einzelner Staaten)^0.5754554 subject:(Internationale Beziehungen)^0.5564111 subject:(Nigeria)^0.4606596 subject:(Regionale Außenpolitik einzelner Staaten)^0.4479771 geo:(Nigeria)^1.0 geo:(Vereinigte Staaten)^0.66411936 geo:(Israel)^0.36490306 subject:(Nigeria)^1.0 subject:(Vereinigte Staaten)^0.850331 subject:(Bilaterale internationale Beziehungen)^0.6000002 subject:(Demokratisierung)^0.39019862 subject:(Außenpolitik einzelner Staaten)^0.30245042 geo:(Nigeria)^1.0 classification:(Außenpolitik)^1.0 subject:(Sowjetunion)^1.0 subject:(Vereinigte Staaten)^0.73333335 subject:(Europäische Gemeinschaften)^0.73333335 subject:(Außenpolitik einzelner Staaten)^0.6666667 subject:(Bundesrepublik Deutschland (1949-1990))^0.53333336 subject:(Außenpolitische Zusammenarbeit)^0.53333336 title:(außenpolitik internationale usa englisch nigeria) außenpolitik internationale usa englisch nigeria

Tabelle 9.3.: Originalanfrage vs. tatsächliche Anfrage SwpEvm1

	Wortlaut Anfrage
Originalanfrage 1:	Wie haben sich die Beziehungen zwischen Nigeria und den USA in den letzten zehn Jahren entwickelt?
Tatsächliche Anfrage 1, generiert im Lauf SwpEvm2:	title:(Beziehungen Nigeria USA)^10 abstract:(Beziehungen Nigeria USA)^10 „Beziehungen“~1 „Relationship“~1 „relations“~1 „relaciones“~1 „Nigeria“~1 „Nigeria“~1 „Nigéria“~1 „Nigeria“~1 „Nigerianer“~1 „Nigerians“~1 „Nigérians“~1 „nigerianos“~1 abstract:(Internationale Beziehungen)^1.0 abstract:(Market Borrowers)^0.44961894 abstract:(AKP-Länder)^0.44961894 abstract:(Severely indebted low-income countries)^0.44961894 abstract:(Englischsprachiges Afrika)^0.44961894 abstract:(Länder mit Schuldendienstproblemen)^0.44961894 subject:(Vereinigte Staaten)^1.0 subject:(Loyalität)^0.9293553 subject:(Disengagement)^0.9293553 subject:(Multilateral)^0.9293553 subject:(Soziale Kontrolle)^0.5808471 subject:(Länder mit Schuldendienstproblemen)^0.5351242 subject:(Vereinigte Staaten)^1.0 subject:(Außenpolitik einzelner Staaten)^0.5516139 subject:(Regionale Außenpolitik einzelner Staaten)^0.43590385 subject:(Internationale Beziehungen)^0.3887316 geo:(Vereinigte Staaten)^1.0 geo:(Nigeria)^0.9718209 geo:(Israel)^0.4103348 geo:(Japan)^0.32219598 geo:(Volksrepublik China)^0.31077272 subject:(Vereinigte Staaten)^1.0 subject:(Nigeria)^0.86106074 subject:(Volksrepublik China)^0.532139 subject:(Japan)^0.48474848 subject:(Bilaterale internationale Beziehungen)^0.44641984 subject:(Außenpolitik einzelner Staaten)^0.3378379 geo:(Volksrepublik China)^1.0 geo:(Lateinamerika)^0.68738633 geo:(Mexiko)^0.4610898 geo:(Afrika (insgesamt))^0.44438052 geo:(Zentralamerika)^0.41343674 geo:(Japan)^0.40242276 classification:(Außenpolitik)^1.0 subject:(Außenpolitik einzelner Staaten)^1.0 subject:(Sowjetunion)^0.75438595 subject:(Vereinigte Staaten)^0.64912283 subject:(Bundesrepublik Deutschland (1949-1990))^0.54385966 subject:(Bilaterale internationale Beziehungen)^0.45614034 subject:(Europäische Gemeinschaften)^0.42105263 title:(außenpolitik internationale usa englisch nigeria) außenpolitik internationale usa englisch nigeria

Tabelle 9.4.: Originalanfrage vs. tatsächliche Anfrage SwpEvm2

	Wortlaut Anfrage
Originalanfrage 1:	Wie haben sich die Beziehungen zwischen Nigeria und den USA in den letzten zehn Jahren entwickelt?
Tatsächliche Anfrage 1, generiert im Lauf SwpEvm3:	title:(Beziehungen Nigeria USA)^10 abstract:(Beziehungen Nigeria USA)^10 „Beziehungen“~1 „Relationship“~1 „relations“~1 „relaciones“~1 „Nigeria“~1 „Nigeria“~1 „Nigéria“~1 „Nigeria“~1 „Nigerianer“~1 „Nigerians“~1 „Nigérians“~1 „nigerianos“~1 abstract:(Internationale Beziehungen) subject:(Vereinigte Staaten) subject:(Loyalität) subject:(Disengagement) subject:(Multilateral) subject:(Vereinigte Staaten) geo:(Vereinigte Staaten) geo:(Nigeria) subject:(Nigeria) geo:(Vereinigte Staaten) classification:(Außenpolitik) subject:(Außenpolitik einzelner Staaten) subject:(Sowjetunion) subject:(Vereinigte Staaten) title:(außenpolitik internationale usa englisch nigeria) außenpolitik internationale usa englisch nigeria

Tabelle 9.5.: Originalanfrage vs. tatsächliche Anfrage SwpEvm3

mehr irrelevante Dokumente auswerten und somit die Ergebnisse verwässern könnte.

SwpEvm2 vs. SwpEvm3

SwpEvm2 unterscheidet sich gegenüber SwpEvm3 in der Veränderung des Schwellenwerts *CutOffScore* von 0,3 (SwpEvm2) auf 0,6 (SwpEvm3). In dem direkten Vergleich zwischen SwpEvm2 und SwpEvm3 wird sich zeigen, ob die Verminderung der Anzahl der eingesetzten Terme durch die Erhöhung des Schwellenwertes bei gleichzeitig höherer Anzahl von untersuchten Dokumenten im Vergleich zu SwpEvm1 die Retrievalqualität steigern kann.

Während durch die (gegenüber SwpEvm1) größere Anzahl von untersuchten Dokumenten gegebenenfalls auch mehr irrelevante und unpassende Deskriptoren durch die statistische Analyse ausgewertet wurden, kann der höhere Score-Schwellenwert ggf. dazu führen, dass die Qualität der zu Anfrage hinzugefügten Metadaten steigt.

Ein Argument für die Verbesserung der Ergebnisqualität kann sein, dass die irrelevanten der 100 untersuchten Dokumente keine einheitliche, inhaltliche Richtung verfolgen dürften, sondern sich in alle thematischen Richtungen vom Cluster der relevanten Dokumente entfernen. Daher ist nicht zu erwarten, dass eine besondere thematische Häufung bei den irrelevanten Dokumenten auftritt. Entsprechend kommt in diesen Dokumenten eine sehr große Anzahl verschiedener Deskriptoren vor, bei denen Mehrfachnennungen unwahrscheinlicher sind als bei den Deskriptoren der relevanten Dokumente. Entsprechend sollte ein höherer Schwellenwert dazu führen, dass der Anspruch an die Häufigkeit und die Bewertung eines Deskriptors deutlich steigt und somit unnütze Deskriptoren eher herausgefiltert werden.

SwpEvm3 vs. SwpEvm1

SwpEvm3 generiert die kürzeste Anfrage der drei Evm-Läufe. Dies liegt an dem hohen Wert für den Parameter *cutOfScore* im EVM-Prozess und die damit verbundene kleinere Anzahl von zurückgegebenen Metadaten. Im Vergleich zu der gegenüber SwpEvm1 erhöhten Anzahl von untersuchten Dokumenten, wird auf diesem Wege eine sehr große Menge von Dokumenten - und damit potentiell auch viele ungeeignete Dokumente - statistisch ausgewertet und unter dem Vielfachen der den Dokumenten angehängten Deskriptoren, die nach Score und Mehrfachnennungen wichtigsten Metadaten extrahiert. Durch die Normalisierung der Score-Werte und den hohen Grenzwert von 0,6 wird nur ein kleiner Anteil aller ausgewerteter Metadaten zurückgegeben. Die vom EVM zurückgegebenen Terme und Phrasen werden nicht mit dem Gewicht verknüpft, was bedeutet, dass alle Deskriptoren mit einem einfachen Gewicht eingehen. Dies ist eine deutliche Vereinheitlichung und in den meisten Fällen auch eine Erhöhung der Gewichtung der Terme. Sowohl in SwpEvm1 als auch in SwpEvm2 gingen mehr Terme mit dynamischen, meist geringeren Gewichten ein, in SwpEvm3 gehen weniger Terme mit einem einheitlich einfachen Gewicht ein.

Für eine Verschlechterung des Retrievalergebnisses spricht die höhere Anzahl von ausgewerteten Dokumenten, da mit einer größeren Anzahl von Dokumenten der Anteil an irrelevantem Ballast üblicherweise überproportional ansteigt.

Für eine Verbesserung der Retrievalergebnisse spricht, dass ein höherer *cutOffScore* ausschließlich die Deskriptoren von sehr hoch bewerteten Dokumenten bzw. die besonders häufig genannten Deskriptoren zurückgeben lässt.

9.3.3. Hinweise zur Wahl der Parameter

Im Rahmen der Entwicklung hat sich während mehrerer Testläufe gezeigt, dass ein Wert von unter 30 Dokumenten für den Paramter *consideredDocs* sich nicht als sinnvoll herausgestellt hat, da auf diese Weise (z.B. bei nur 10 untersuchten Dokumenten) die relevanten Deskriptoren nicht deutlich genug durch Mehrfachnennungen ihren Score steigern konnten. Werden beispielsweise semantisch verwandte Deskriptoren für die Beschreibung von Dokumenten verwendet, so kam in einer kleinen Menge von untersuchten Dokumenten für die Mehrzahl der untersuchten Anfragen keine Häufung der relevanten Deskriptoren zustande. Die andere Vergleichsgröße für diesen Parameter, 100, wurde gewählt, um einen möglichst eindeutigen Unterschied festzustellen. Die Zahl wurde bewusst nicht noch höher angesetzt, da zu befürchten war, dass Dokumente, die einen noch schlechteren Score erzielten, keinen sehr wertvollen Beitrag mehr liefern konnten. Außerdem sinkt das Gewicht der ergänzten Deskriptoren bei jedem

	SwpBase1	SwpBase 2	SwpEvm1	SwpEvm2	SwpEvm 3
Eingesetzte Module:					
Discriminator	-	-	X	X	X
Translator	-	-	X	X	X
BRF	-	-	X	X	X
EVM	-	-	X	X	X
Eingesetzte Anfragen:					
EnteredQuery	X	X	-	-	-
StoppedQuery	-	-	X	X	X
TranslatedQuery	-	-	X	X	X
RelevantTerms	-	-	X	X	X
EvmQuery	-	-	X	X	X
Parameter des EVM:					
consideredDocs	-	-	30	100	100
numberOfReturned	-	-	6	6	6
cutOffScore	-	-	0,3	0,3	0,6
Scoreberechnung	-	-	normalisiert	normalisiert	normalisiert
EvmQuery	-	-	X	X	X
Gewichtung der Anfragepassagen:					
EnteredQuery	1	1	-	-	-
StoppedQuery	-	-	10	10	10
TranslatedQuery	-	-	1	1	1
RelevantTerms	-	-	1	1	1
EvmQuery	-	-	Score	Score	1
Bezug der Anfrage auf folgende Indexfelder:					
title	X	X	X	X	X
abstract	X	X	X	X	X
subject	-	X	X	X	X
classification	-	X	X	X	X
geo	-	-	X	X	X

Tabelle 9.6.: Überblick über alle Parameter der Evaluierungsläufe

weiteren Schritt auf der Ergebnisliste, entsprechend hätten die Deskriptoren der unteren Dokumente ohnehin nur ein sehr geringes Gewicht gehabt. Eine interessante, weitere Option, die in diesem Rahmen nicht untersucht wurde, ist die relative Anpassung der Anzahl der untersuchten Dokumente an die Gesamtanzahl der gefundenen Dokumente. Die genauere Untersuchung einer solchen Anpassung ist - neben der genaueren Analyse einer Vielzahl von verschiedenen weiteren Konfigurationsmöglichkeiten - an dieser Stelle nicht möglich, ohne den Rahmen dieser Arbeit zu sprengen.

Die Werte 0,3 und 0,6 wurden für den Parameter *cutOffScore* festgelegt, um einen deutlichen Unterschied zwischen einer längeren (und damit qualitativ möglicherweise schwächeren) und einer kürzeren (und damit qualitativ möglicherweise konkreteren) Anfrage zu simulieren. Die gewählte Anzahl von sechs Termen für *numberOfReturned* sollte in jedem Fall die Anzahl der durch jedes einzelne EVM-Element ergänzten Deskriptoren limitieren. Diese Anzahl schein geeignet, da die durchschnittliche Anzahl

von thematischen Deskriptoren pro Dokument bei 12,2 liegt. Entsprechend scheint sich der Inhalt eines Dokuments im Durchschnitt mit rund zwölf Deskriptoren geeignet beschreiben zu lassen. Da allerdings nicht zu erwarten ist, dass tatsächlich die zwölf am höchsten bewerteten Deskriptoren eines Kaskadenelements auch geeignet sind, wurde diese Anzahl halbiert. Dieser Parameter kommt allerdings nur dann zum Tragen, wenn tatsächlich mehr als sechs Deskriptoren hinzugefügt werden könnten. Im Falle von geopolitischen Informationen ist dies nicht zu erwarten, da pro Dokument im Schnitt nur 1,2 Deskriptoren vergeben wurden. Hier ist zu erwarten, dass die Hauptaufgabe, die relevanten Deskriptoren auszuwählen, durch die Funktion des *cutOffScore* erreicht wird und selten eine größere Anzahl von Deskriptoren empfohlen werden.

Durch maximal sechs zurückgegebene Deskriptoren wird die Anzahl der durch das EVM ergänzten Terme und Phrasen auf maximal 54 (bei 9 Kaskadenelementen) begrenzt. Diese maximal 54 Deskriptoren werden auf insgesamt drei Indexfelder (*subject*, *geo* und *classification*) gerichtet. Da mehrere Kaskadenelemente Deskriptoren aus dem selben Indexfeld extrahieren und Ihre Ergebnisse wieder entsprechend ausrichten, ist zu erwarten, dass sich diese potentiell sehr große Anzahl von Deskriptoren durch Doppelnennungen weiter verkürzt und sich auch hier die Gewichtungen von doppelt genannten Deskriptoren aufaddieren. Da jedoch davon ausgegangen wird, dass die ursprüngliche Anfrage nach wie vor die beste Grundlage für einen Retrievalvorgang bietet, soll diese Passage der Anfrage auch deutlich betont werden. In Probeläufen während der Entwicklung zeigte sich, dass Deskriptoren durch Doppelnennungen bis zu fünffache Gewichtungen erreichten. In dem vorliegenden System erhält die reduzierte Ausgangsanfrage daher eine zehnfache Gewichtung, um diese Passage besonders zu betonen. Es wird erwartet, dass die originale Anfrage nach wie vor den besten Ansatz für das Auffinden von relevanten Dokumenten darstellt, somit sollen die ergänzten Passagen die Anfrage verbessern und nicht gänzlich verändern.

Die Änderung der Anfrageformulierung in SwpEvm3 scheint im Nachhinein die Vergleichbarkeit bei der Untersuchung der Veränderung des *cutOffScore* zu untergraben, eine gänzlich stringente Formulierung hätte die Ergebnisse ggf. eindeutiger werden lassen können. Allerdings sei darauf hingewiesen, dass die Veränderung der Anfrageformulierung durch das Entfernen der Gewichte an den Termen ohne gravierende Auswirkung bleiben dürfte, da die Gewichte der hinzugefügten Deskriptoren ohnehin im, in einer Stichprobe gemessenen, Mittel nur um 0,083 von einem Gewicht von 1 abweichen.

9.3.4. Moduleinsatz in den Evaluierungsläufen

Während die SwpEvm-Läufe alle Module einsetzen, werden bei den SwpBase-Läufen auf Wunsch der SWP keinerlei zusätzliche Retrievalmodule eingesetzt. Insgesamt lässt sich also an den Unterschieden zwischen SwpEvm- und SwpBase-Läufen ablesen, welcher Unterschied zwischen einem IR-System ohne jegliche und einem IR-System mit allen vorgestellten Modulen entsteht. Da sich die Arbeit im Speziellen auf das Entry Vocabulary Modul bezieht und damit vorallem die Parameter zur Einstellung dieses Moduls untersucht werden sollen, wurden die drei Läufe mit Modulen entsprechend der Erforschung der Parameterisierung und nicht dem Unterschied Blind Relevance Feedback/Translator/Entry Vocabulary Module gewidmet. Insgesamt wurde die Wahl der Gewichtungen und Anzahl der hinzugefügten Terme dem Schwerpunkt der Analyse, dem EVM, angepasst. In einem separaten, ergänzenden Evaluierungsverfahren wird allerdings die Leistungsfähigkeit der einzelnen Module durch die zusätzlichen Evaluierungsläufe SwpEvm_only, SwpBrf_only und SwpTranslator_only annäherungsweise untersucht. Die Ergebnisse finden sich in Abschnitt 10.1.4 ab Seite 82.

9.4. Evaluierung der mehrsprachigen Retrievalleistung

Während die Evaluierung der Mehrsprachigkeit eines Retrievalsystems üblicherweise durch eine normale Relevanzbewertung von Ergebnissen erzielt wird, die aus einem Retrievalvorgang entstanden, bei denen Anfragesprache und Zielsprache differieren, lagen im Rahmen dieses Projekts keine fremdsprachlichen Anfragen vor. Vielmehr ist zu erwarten, dass das System anhand von deutschen Anfragen auf alle verzeichneten, und damit auch die nicht-deutschsprachigen, Dokumente durchsucht wird.

Entsprechend wurde eine Evaluierungsmethode entwickelt, bei dem die vorliegende Evaluierungssoftware dazu verwendet wird, alleine die Anzahl der gefundenen nicht deutschsprachigen Dokumente zu vergleichen. Die Ergebnisse hierzu sind in Abschnitt 10.1.3 ab Seite 80 zu finden.

Wie bereits in Kapitel 4.1.5 auf Seite 31 beschrieben wurde, verzeichnet der Datenbestand des Fachinformationsverbundes nur zu einem relativ kleinen Teil Dokumente in deutscher Sprache, entsprechend sollte auch ein auf dieser Datengrundlage entwickeltes System und vor allem das Entry Vocabulary Modul durch die Nutzung des kontrollierten Vokabulars eine deutliche Verbesserung im mehrsprachigen Retrievalergebnis erreichen.

Zu Evaluierungszwecken der mehrsprachigen Retrievalleistung wurden die vorlie-

genden Relevanzbewertungen verwendet. Es wurden alle in den Relevanzbewertungen als relevant bezeichneten Dokumente auf Ihre Sprache hin untersucht und bei Fremdsprachigkeit ein entsprechender Vermerk ergänzt. Die Auswertung des Recall-Werts geschieht daraufhin mit Hilfe des selben Evaluierungstools, das auch für die allgemeinen Auswertungen herangezogen wurde. Ziel der Auswertung ist zu erfahren, wie viele der für relevant befundenen Dokumente in einer anderen Sprache als der deutschen verfasst wurden.

10. Auswertung der Evaluierung

Nachdem im vorhergegangenen Kapitel die Vorbereitung der Evaluierungsläufe detailliert beschrieben wurde, sollen an dieser Stelle die Ergebnisse der Evaluierung in diversen Ansätzen präsentiert werden.

Abschnitt 10.1 gibt einen Gesamtüberblick über alle Evaluierungsläufe, die auf die SWP-Datenbasis angewendet wurden. Diese Analyse gliedert sich in eine zusammenfassende Auswertung der Ergebnisse über alle Anfragen in Abschnitt 10.1.1 und eine detailliertere Auswertung der Ergebnisse in Abschnitt 10.1.2, die die Retrievalleistung der einzelnen Läufe pro Anfrage untersucht.

Abschnitt 10.2 dokumentiert und analysiert die Evaluierung auf den GIRT- Datenbestand. Hierbei wird in Abschnitt 10.2.1 ein kurzer Einblick in GIRT3 gegeben, in Abschnitt 10.2.2 die Anpassungen des für FIV entwickelten Systems bei der Anwendung auf GIRT beschrieben und in Abschnitt 10.2.3 die Ergebnisse der Evaluierung für GIRT präsentiert.

Abschnitt 10.3 fasst die wichtigsten Ergebnisse zusammen und versucht, die zentralen Fragen aus der Einleitung zu beantworten.

10.1. Ergebnisse für die Datenbasis des Fachinformationsverbundes

Die Ergebnisse der Evaluierung zeigen, dass die Leistung des Retrievalsystems durch die zusätzlich eingesetzten Module sowohl im Recall (deutlich) als auch in der Precision (ansatzweise) gesteigert werden konnte (vgl. Tabelle 10.1 und Abbildung 10.1). Gleichzeitig fällt auf, dass alle Evaluierungsläufe noch keine besonders guten Ergebnisse erzielen konnten, da auch nicht die weiterentwickelten Läufe eine höhere, durchschnittliche Precision als 0,22 erreichten. Es zeigt sich jedoch über alle fünf Evaluierungsläufe ein deutlicher Precision-Anstieg auf knapp 0,6 (SwpBase1 und SwpEvm3) bzw. 0,7 (SwpEvm2 und SwpEvm1) unter den ersten 10% des Recalls. In den folgenden Unterkapiteln sollen zunächst die einzelnen Läufe miteinander verglichen und schließlich die Gruppen SwpBase und SwpEvm einander gegenübergestellt werden.

	SwpBase1	SwpBase2	SwpEvm1	SwpEvm2	SwpEvm3
Retrieved	5000	5000	5000	5000	5000
Relevant	2039	2039	2039	2039	2039
Rel-Ret	717	965	1226	1182	1211
0	0,5815	0,6719	0,6787	0,6803	0,5801
0,1	0,2596	0,3735	0,4085	0,3860	0,4008
0,2	0,1945	0,2971	0,3622	0,3446	0,3281
0,3	0,1541	0,2295	0,3252	0,3079	0,3073
0,4	0,1106	0,2065	0,2932	0,2665	0,2806
0,5	0,0671	0,1623	0,2349	0,2084	0,2672
0,6	0,0413	0,1420	0,1784	0,1498	0,1639
0,7	0,0057	0,0836	0,1213	0,1164	0,1056
0,8	0,0000	0,0432	0,0330	0,0461	0,0240
0,9	0,0000	0,0082	0,0123	0,0000	0,0032
1	0,0000	0,0025	0,0017	0,0000	0,0031
Durchschnitt	0,0985	0,1752	0,2138	0,1980	0,2046

Tabelle 10.1.: Precision-Ergebnisse über alle Läufe

10.1.1. Analyse der Evaluierungsläufe über alle Anfragen

SwpBase1 vs. SwpBase2

Die Recall-Ergebnisse der Evaluierungsläufe SwpBase1 und SwpBase2 unterscheiden sich deutlich voneinander. Die Anzahl der gefundenen, relevanten Dokumente ist um 35% gestiegen, während sich die durchschnittliche Precision um fast 80% gesteigert hat. Die beste Precision-Leistung ergibt sich bei beiden Basisläufen unter den ersten 10% der relevanten Dokumente. Hier erzielen beide Varianten, gemessen am übrigen Verlauf des Recall-Precision-Diagramm, überproportionale Werte von 0,582 (SwpBase1) und 0,672 (SwpBase2).

SwpEvm1 vs. SwpEvm2 vs. SwpEvm3

Alle drei Evm-Läufe liegen im Recall nahe beieinander, es ergibt sich eine Differenz von nur 42 Dokumenten (3,5% gemessen an SwpEvm2), die den leistungsfähigsten Lauf SwpEvm1 vom schwächsten Evm-Lauf SwpEvm2 trennen. Die durchschnittliche Precision liegt bei allen drei Läufen um 0,200, auch hier bestätigt sich, dass SwpEvm1 der leistungsfähigste Lauf ist mit 0,214 und SwpEvm2 der schwächste mit 0,198. In der höchsten Recall-Stufe kann SwpEvm3 allerdings nicht die gleiche Precision wie die anderen Evm-Läufe und auch SwpBase2 erreichen. Insgesamt ergeben sich allerdings, im Gegensatz zu den Unterschieden zwischen den Base-Läufen, keine größeren Unterschiede zwischen der Retrievalqualität der einzelnen Läufe. SwpEvm3 liegt sowohl in Recall (1211) als auch in der Precision (0,205) zwischen den beiden anderen Evm-Läufen, liegt allerdings näher beim stärkeren Lauf SwpEvm1 im Recall.

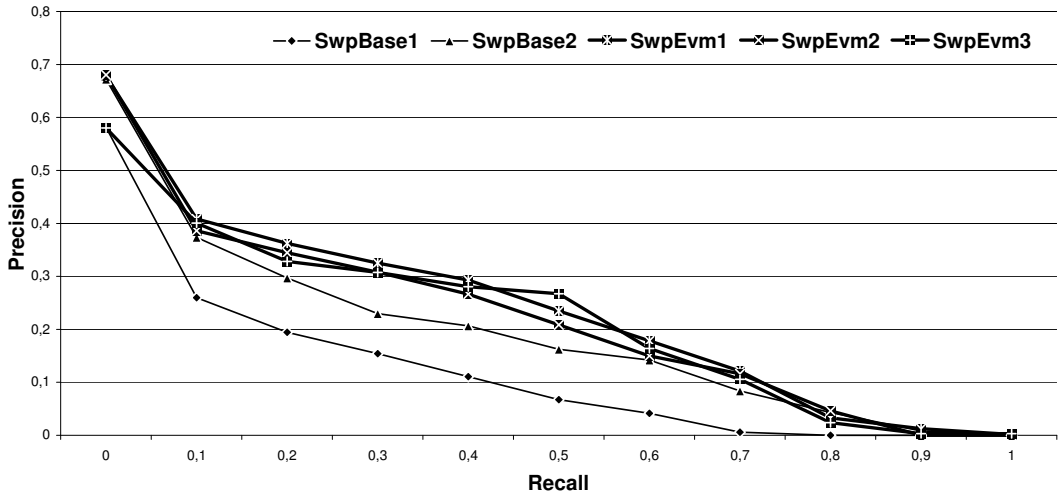

Abbildung 10.1.: Precision-Ergebnisse über alle Läufe

SwpBase vs. SwpEvm

Die Steigerung der Leistung lässt sich am deutlichsten in den Recall-Ergebnissen ablesen. Während die Evaluierungsläufe SwpBase1 und SwpBase2 nur 717 bzw. 965 relevante Dokumente gefunden haben, konnten die SwpEvm-Läufe zwischen 1182 und 1226 Dokumente abbilden. Dies entspricht einer Steigerung von 43,4% des Durchschnitts der SwpBase-Läufe (841) gegenüber dem Durchschnitt der SwpEvm-Läufe (1206,3).

Die Precision-Werte unterscheiden sich nicht im gleichen Maße wie die Recall-Werte, SwpBase1 sei hiervon ausgenommen. Alle drei SwpEvm-Läufe und auch SwpBase2 erreichen eine Precision von rund 0,200 wobei SwpBase2 mit 0,174 klar der schwächste Lauf ist. Allerdings zeigt sich hierdurch, dass alleine durch die Anwendung der Anfragenterme auf die Metadatenfelder *subject* und *classification* ein großer Fortschritt gegenüber SpwBase1 in der Precision sowie im Recall erzielt werden konnte. Innerhalb der ersten 10% der relevanten Dokumente ergibt sich für SwpBase2 sogar ein deutlich stärkeres Retrievalergebnis als für SwpEvm3.

10.1.2. Analyse der Evaluierungsläufe über einzelne Anfragen

Tabelle 10.2 zeigt eine Auflistung der Anfragen nach Precision pro Lauf. In dieser Darstellung lässt sich ablesen, welche Anfragen im Kontext eines Laufs vergleichsweise erfolgreich waren und welche Anfragen im selben Lauf schlechtere Ergebnisse erzielt

Lauf		SwpBase1		SwpBase2		SwpEvm1		SwpEvm2		SwpEvm3
Precision	23	0,0106	10	0,0025	20	0,0054	20	0,0042	10	0,0006
nach	15	0,0115	23	0,0036	11	0,0296	23	0,0223	20	0,0011
Anfrage	5	0,0152	14	0,0344	23	0,0391	11	0,0275	11	0,0102
sortiert	11	0,0153	25	0,0420	2	0,0472	21	0,0361	23	0,0204
	10	0,0233	17	0,0420	17	0,0521	2	0,0518	25	0,0567
	20	0,0253	18	0,0511	25	0,0567	25	0,0558	21	0,0774
	24	0,0263	24	0,0558	9	0,0908	15	0,0627	8	0,0800
	3	0,0266	11	0,0614	15	0,0989	4	0,0829	13	0,0803
	25	0,0307	1	0,0714	8	0,1041	19	0,0889	4	0,0979
	17	0,0344	15	0,0739	19	0,1088	8	0,0920	19	0,1040
	7	0,0596	13	0,0829	4	0,1100	9	0,1073	7	0,1141
	4	0,0599	6	0,0838	3	0,1287	17	0,1293	2	0,1328
	14	0,0599	3	0,0841	7	0,1401	3	0,1317	14	0,1640
	1	0,0635	4	0,0887	13	0,1438	5	0,1325	5	0,1848
	6	0,0712	20	0,1131	18	0,1440	1	0,1688	17	0,2061
	12	0,0751	12	0,1240	5	0,1681	14	0,2013	9	0,2170
	13	0,0836	7	0,1543	14	0,2107	24	0,2027	24	0,2389
	9	0,0996	2	0,1674	21	0,3246	7	0,2093	18	0,2506
	19	0,1021	9	0,1721	24	0,3264	13	0,2313	15	0,2519
	8	0,1040	19	0,1931	10	0,3416	18	0,2716	1	0,2668
	21	0,1309	5	0,2182	1	0,3825	10	0,2784	3	0,3248
	18	0,1657	8	0,3762	12	0,4153	6	0,4429	6	0,3392
	2	0,1674	21	0,5672	6	0,4362	12	0,5156	12	0,4039
	22	0,3843	16	0,7437	22	0,6834	22	0,6350	22	0,7299
	16	0,6168	22	0,7735	16	0,7573	16	0,7687	16	0,7626
Durchschnitt		0,0985		0,1752		0,2138		0,1980		0,2046

Tabelle 10.2.: Precision-Ergebnisse nach Anfrage über alle Evaluierungsläufe

haben. Darüber hinaus lässt sich ablesen, wie die Leistung der Anfragen sich über alle Läufe im Kontext der anderen Anfragen verändert hat. Möglicherweise lässt sich ein Schema ablesen, das zeigt, dass bestimmte Anfragetypen für einige Läufe besser geeignet sind als für andere. Da für eine solche Analyse eine große Anzahl von Evaluierungsanfragen nötig ist, um repräsentativ zu sein, kann diese Analyse im besten Fall einen Trend unter den im Rahmen dieses Projekts verwendeten Anfragen beschreiben, jedoch keine exakten, verlässlichen Regeln formulieren.

Verteilung der Anfrageleistung und Homogenität

Zur besseren Einschätzung der Leistung der einzelnen Anfragen wurden in Tabelle 10.2 zusätzliche, vertikale Linien eingezogen, um das gesamte Feld der Anfragen zu gruppieren. Die (von oben nach unten gezählte) erste Linie grenzt die Anfragen mit einem schlechteren Ergebnis als 30% der durchschnittlichen Precision ab. Die zweite Linie zeigt den mittleren Durchschnitt. Die dritte Linie grenzt die überdurchschnittlichen Anfragen von den herausragenden Anfragen ab, die über 150% des Durchschnitts erreicht haben. Bei SwpEvm1 fallen die zweite und die dritte Linie auf die gleiche Stelle

zwischen Anfrage 14 und 21.

Zunächst fällt auf, dass der Anteil der Anfragen, die mit ihrem Precision-Wert über dem Mittelwert des gesamten Evaluierungslaufs liegen, von 28% bei SwpBase1/2 auf 44% bei SwpEvm3 gesteigert werden konnte. Das bedeutet, dass die Ergebnisqualität über einen Großteil der Anfragen bei den Evaluierungsläufen stetig geworden ist und durchschnittlich mehr Anfragen von einer gesteigerten Retrievalleistung profitieren. Diese Feststellung wird zusätzlich davon unterstützt, dass der Anteil der Anfragen, die mit weniger als einem Drittel der durchschnittlichen Precision am schlechtesten abschneiden von 36% bei SwpBase1 auf 20% bei SwpEvm3 sinkt. Insgesamt erzielt das System also im Evaluierungslauf SwpEvm3 eine gewisse Verbesserung in der Homogenität der Retrievalleistung nach Precision.

Zusätzlich ist zu beachten, dass der Anteil der Anfragen, die eine 150% bessere Leistung als der Durchschnitt erbrachten, bei SwpEvm1 mit 32% im Vergleich zu den anderen Läufen (16-20%) besonders hoch liegt. Alle 8 Anfragen, die im Rahmen der Precision über dem Durchschnitt liegen, haben mehr als 150% der Durchschnittsleistung erreicht. Damit verfügt SwpEvm1 über eine große Gruppe an sehr erfolgreichen Anfragen und ein kleineres Mittelfeld (zwischen 150% und 30% der durchschnittlichen Precision). SwpEvm2 hat dagegen nur wenige herrausragende Anfragen, wobei diese vier Anfragen bei nahezu allen Läufen besonders erfolgreich waren.

Entwicklung der Leistung der einzelnen Anfragen

Interessant ist jedoch auch, die Entwicklung der Precision einzelner Anfragen über alle Läufe zu betrachten: In diesem Rahmen werden zunächst die überdurchschnittlich abschneidenden Anfragen für SwpBase1 und SwpEvm1 und darauf folgend die jeweils vier schlechtesten Anfragen für die selben beiden Läufe untersucht.

Zunächst soll jedoch festgehalten werden, dass die Anfragen 16 und 22 in allen fünf Evaluierungsläufen an Position 1 und 2 mit den entsprechenden besten Precision-Werten und mit deutlichem Abstand zu den weiteren Anfragen stehen. Diese beiden Anfragen werden für diese Analyse ausgeklammert, da hier offensichtlich keine Veränderung zwischen den Läufen festzustellen ist. Eine nähere Betrachtung dieser Anfragen wird im folgenden Unterkapitel beschrieben.

Bei der Untersuchung der Anfragen mit den überdurchschnittlichen Precision-Werten von SwpBase1 fällt auf, dass keine der fünf Anfragen in einem der Evm-Läufe höher gelistet wird, die Entwicklung verläuft über ein weitgehend ähnliches Abschneiden mit durchweg deutlich verbesserten Precision-Werten bei SwpBase2 meist auf einen hinteren Platz bei der Evm-Gruppe mit deutlich schlechteren Precision-Werten als bei SwpBase2. Allein Anfrage 18 eignet sich vergleichsweise schlecht für SwpBase2 und

hat größeren Erfolg bei den Evm-Läufen.

Die Anfragen mit einem überdurchschnittlichen Precision-Ergebnis bei SwpEvm1 hatten durchweg nur unterdurchschnittliche Ergebnisse in SwpBase1 und sogar noch deutlich schlechtere Platzierungen (bei geringfügig besseren Precision-Werten) in SwpBase2. Die durchschnittliche Precision konnte bei diesen Läufen von 0,052 (SwpBase1) über 0,068 (SwpBase2) auf 0,3804 (SwpEvm1) gesteigert werden. Die überdurchschnittlichen Anfragen für SwpEvm1 erzielen auch bei den anderen Evm-Läufen weitestgehend überdurchschnittliche Ergebnisse, einzig der drastische Unterschied der Leistung der Anfrage 10 zwischen den Läufen SwpEvm1/SwpEvm2 (positiv) und den Base-Läufen sowie SwpEvm3 (sehr negativ) ist weiter zu untersuchen.

Die Entwicklung der vier schlechtesten Anfragen für SwpBase1 zeichnet kein einheitliches Bild: Die Anfragen 11 und 23 bleiben unterdurchschnittlich, während die Anfrage 15 im Lauf SwpEvm3 eine Verbesserung der Precision von 0,012 auf 0,2519 erreicht. Anfrage 5 erreicht in SwpBase2 ihr bestes Ergebnis mit einer Verbesserung um 0,2030.

Die Entwicklung der vier schlechtesten Anfragen für SwpEvm1 ergibt über die anderen Evm-Läufe ebenfalls unterdurchschnittliche Ergebnisse. Anfrage 2 war jedoch die drittbeste Anfrage für SwpBase1. Sowohl Anfrage 20 als auch Anfrage 2 sind im Rahmen von SwpBase2 erfolgreicher gewesen als bei den Evm-Läufen.

Prüfung ausgewählter Suchanfragen

Während aus Platzgründen nicht alle tatsächlich eingesetzten Anfragen über alle Evaluierungsläufe untersucht werden können, soll in diesem Unterkapitel in mehreren exemplarischen Fällen nachvollzogen werden, wie die bemerkenswertesten Fälle von Unterschieden in der Precision einer Anfrage über mehrere Läufe zustande kamen. Alle vorgestellten Beispiel lassen sich auch in den Output-Dateien auf der beiliegenden CD nachvollziehen.

> 16. Welches sind die zentralen Herausforderungen für die Europäische Sicherheits- und Verteidigungspolitik?
>
> 22. Wie hat sich die Gemeinsame Außen- und Sicherheitspolitik der EU in den letzten Jahren entwickelt?

Die Anfragen 16 und 22 waren über alle Evaluierungsläufe im herausragenden Maße erfolgreich. Eine Untersuchung der Output-Dateien, die für jeden Evaluierungslauf die eingesetzten Suchanfragen protokollieren, ergibt, dass Anfrage 16 bei den EVM-Läufen überdurchschnittlich viele Ergänzungen aus dem Translator-Modul (52 Phrasen) erhalten hat. Diese Beiträge zur Anfrage sind zu einem Großteil passend, umfassen aber

auch Übersetzungen der Begriffe „Wirtschaftliche Sicherheit“ und „Technische Sicherheit“, die nicht für den Inhalt der Anfrage zentral sind. Die 32 Beiträge aus dem EVM sind allerdings durchweg relevant und hochwertig, entsprechend kann davon ausgegangen werden, dass diese Terme das Ergebnis der Anfrage positiv beeinflusst haben. Anfrage 22 wurden durch das Translator Modul 16 Phrasen hinzugefügt, die überwiegend geeignet waren, während durch das EVM 28 Deskriptoren ergänzt wurden, die ebenfalls nahezu alle zur Erweiterung geeignet waren. Da diese beiden Anfragen jedoch auch in SwpBase1 und SwpBase2 überaus erfolgreich waren, haben die ergänzten Passagen offensichtlich keine weitere Verbesserung realisieren oder Verschlechterung verursachen können.

Frage 16 und 22 sind sowohl thematisch als auch geopolitisch nahezu identisch und haben beide den Charakter einer sehr offenen Anfrage, entsprechend wurde vermutlich sehr viel gefundenes Material als relevant bewertet. Zusätzlich dazu ist zu erwarten, dass dieses Thema im Datenbestand umfassend behandelt wird und entsprechend eine gute Chance bestand, hervorragende Ergebnisse zu erzielen. Bemerkenswert bei diesen beiden Anfragen ist, dass sowohl die Base-Läufe als auch die Evm-Läufe nahezu die selben Werte der Precision erreicht haben. Es ist zu erwarten, dass vielen Dokumente zum Thema mit Zusammenfassungen ausgestattet sind, da sonst SwpBase1 wahrscheinlich nicht im gleichen Maße positiv hätte abschneiden können.

> 11. Wie gestaltet die erweiterte Europäische Union ihre Beziehungen gegenüber den europäischen Staaten der ehemaligen Sowjetunion und gegenüber dem Westlichen Balkan?

> 23. Wie gestaltet sich die Zukunft des mediterranen Raums aus europäischer Sicht?

Anfrage 23 ist die einzige Anfrage, die konstant über alle Evaluierungsläufe unter den letzten vier Anfragen aufgelistet wird. Die maximale Precision für diese Anfrage liegt bei SwpEvm1 mit einem sehr schlechten Wert von 0,0391. Ähnlich schlecht schnitt nur Anfrage 11 über alle Läufe ab. Beide Anfragen wurden nach der in Abbildung 9.1 umgesetzten Kategorisierung mit vagen Themenangaben und der Nennung von überregionalen Räumen verzeichnet. Die durch das EVM ergänzten Terme und Phrasen bei Anfrage 11 beziehen sich zu weiten Teilen sehr einseitig auf die EU und EU-interne Themenfelder wie das Währungssystem, das Zentralbanksystem und andere ungeeignete Themengebiete. Insgesamt ließ sich offensichtlich nicht durch die ergänzten Terme der thematische Komplex der Beziehungen zwischen den beiden, in

der Anfrage angesprochenen Parteien, klar abbilden und damit die Leistungsfähigkeit der Frage verbessern.

Anfrage 23 hingegen ist ebenfalls ein gutes Beispiel für eine offene und wenig konkretisierte Frage. Das EVM ergänzt neben nützlichen Passagen wie „Entwicklungsperspektiven und -tendenz“ allerdings auch einen großen Anteil an Deskriptoren, die nicht relevant sind (so zum Beispiel „Lateinamerika“ oder „Vereinigte Staaten“) oder aber zu generalistisch beschreiben (wie beispielsweise „Afrika“ oder „Zeit“) und im weiteren Verlauf der EVM-Kaskade das Ergebnis verschlechtern.

> 10. Wie verhält sich China gegenüber dem nordkoreanischen Atomkonflikt?

> 15. Welche Haltung nehmen die britischen Parteien gegenüber der EU ein?

Die Ergebnisse von Anfrage 10 über alle fünf Evaluierungsläufe sind sehr unterschiedlich. Bei den Base-Läufen schneidet die Anfrage vergleichsweise schlecht ab, während sie bei SwpEvm1 und SwpEvm2 überdurchschnittliche Ergebnisse erzielt, dann in SwpEvm3 als schlechteste Anfrage wieder einbricht. Die Anwendung der Anfragenterme auf die Metadatenfelder in SwpBase2 hat in diesem Fall eher geschadet als geholfen. Bei SwpEvm3 wurde hingegen durch den hohen Cut-Off von 0,6 die erste Nennung von Nordkorea (Score von 0,54) knapp verfehlt, entsprechend beziehen sich die ergänzten Metadaten bei SwpEvm3 ausschließlich auf China bzw. offene Klassifikationsangaben wie „Internationale Wirtschaft/Außenpolitik“. In SwpEvm1 und SwpEvm2 wird Nordkorea genannt und später durch die Kaskade zusätzlich verstärkt (zunächst hat Nordkorea nur ein Gewicht von den oben genannten 0,54, später wird Nordkorea noch viermal genannt (Score-Summe von mehr als 3,0).

Anfrage 15 hat dagegen sehr gute Ergebnisse bei SwpEvm3 gezeigt, jedoch bei allen anderen Evaluierungsläufen sehr schlecht abgeschnitten. Bei den Evm-Läufen wurde diese Anfrage deutlich überdurchschnittlich mit Ergänzungen aus dem Translator Modul erweitert. Darunter sind diverse Übersetzungen von Bezeichnungen für Parteien wie „Liberale Partei“ oder „Republikanische Partei“. Das schlechte Abschneiden der Anfrage in den beiden anderen Evm-Läufen belegt allerdings, dass der Translator keinen maßgeblichen Einfluss auf das Ergebnis bei SwpEvm3 hatte. Der hohe Schwellenwert bei SwpEvm3 hat bewirkt, dass viele Deskriptoren, die irrelevant für die Frage waren (Beispielsweise „Namibia“ (0,34), „Argentinien“ (0,42) oder „Vereinigte Staaten“ (0,34)), nicht für die endgültige Anfrage verwendet wurden. Besonders im Hinblick

auf klar bilateral bezogene Fragen lässt sich darauf schließen, dass eine stärkere Begrenzung der berücksichtigten Geo-Deskriptoren hilfreich sein kann, da im Bestfall nur die zwei am höchsten bewerteten Ergebnisse relevant wären (dies lässt sich außerdem am Fall der Anfrage 1 belegen, in der „Israel“ als Deskriptor durch das EVM mit eingebracht wird).

Anwendung der Ergebnisse auf das entwickelte Anfragenschema

Bei der Kombination der Kategorisierung der Anfragen nach verschiedenen Graden der Konkretisierung (vgl. Abbildung 9.1 auf Seite 59) mit der Ergebnisauflistung in Tabelle 10.2 auf Seite 74 kann eine Übersicht entsprechend Tabelle 10.3 erstellt werden, die die im jeweiligen Bezugssystem erfolgreichste Umgebung für jede einzelne Frage abbildet. In vielen Kategorien kommt es zu Mehrfachnennungen, da mehrere Anfragen in diesen Kategorien ausgewertet wurden und somit unterschiedliche Ergebnisse protokolliert werden müssen.

Es zeigt sich, dass die Anfragen mit vagen Themenangaben in den Evm-Läufen (vor allem SwpEvm1 und SwpEvm3) im Vergleich zu anderen Anfragen besser abgeschnitten haben. Dagegen lässt sich festhalten, dass sowohl bei präzisen, als auch bei oberbegriffsartigen Themenangaben in Anfragen die SwpBase-Gruppe diese im Vergleich zu den anderen Anfragen besser beantworten konnte als die anderen Läufe. In diesen Kategorien lässt sich ebenfalls feststellen, dass hier unter den Evm-Läufen hauptsächlich SwpEvm2 genannt wird.

Darüber hinaus kann festgehalten werden, dass vier der fünf stärksten Anfragen (nach Anfrage 22 und 16) von SwpBase1 keine geopolitischen Anhaltspunkte enthalten. Die Precision konnte sich bei drei der vier Anfragen zu SwpBase2 hin steigern, fiel dann aber deutlich ab. Offensichtlich werden die Ergebnisse des EVM-Systems negativ davon beeinflusst, wenn keine geopolitischen Angaben in der Anfrage gemacht werden da im Rahmen der Ergänzung von Geoinformationen meist ungeeignete Passagen der

Präzise Themenangabe	Base2, Base2, Evm1	Evm2		Evm2	Base1, Base2, Base2, Evm2
Oberbegriffsartige Themenangabe	Base2, Base2, Base1	Evm1	Base2, Evm2, Base1	Evm2, Base2, Evm3	
Vage Themenangabe		Base2, Evm3	Evm1	Evm3	Evm1, Evm3
Geopolitische Angabe	Kein Geo	überreg.	Region	Staatenb.	Staaten

Tabelle 10.3.: Relativ beste Eignung des Evaluierungslaufs nach Anfragetypen

Präzise Themenangabe					
SwpBase1:	0,0871	0,0596		0,0751	0,0789
SwpBase2:	0,3331	0,1543		0,1240	0,1138
SwpEvm1:	0,2517	0,1401		0,4153	0,1212
SwpEvm2:	0,1103	0,2093		0,5156	0,1104
SwpEvm3:	0,1321	0,1141		0,4039	0,0879
Oberbegriffsartige Themenangabe					
SwpBase1:	0,0943	0,0712	0,0581	0,2607	
SwpBase2:	0,1541	0,0838	0,0712	0,4509	
SwpEvm1:	0,1403	0,4362	0,1035	0,3780	
SwpEvm2:	0,1643	0,4429	0,1233	0,3114	
SwpEvm3:	0,1798	0,3392	0,0783	0,3487	
Vage Themenangabe					
SwpBase1:		0,0129	0,0599	0,0266	0,0375
SwpBase2:		0,0325	0,0344	0,0841	0,0726
SwpEvm1:		0,0344	0,2107	0,1287	0,2407
SwpEvm2:		0,0249	0,2013	0,1317	0,1157
SwpEvm3:		0,0153	0,1640	0,3248	0,2594
Geopolitische Angabe	Kein Geo	überreg.	Region	Staatenb.	Staaten

Tabelle 10.4.: Durchschnittliche Precision-Ergebnisse über alle Läufe nach Kategorie

Anfrage hinzugefügt werden.

Eine weitere Möglichkeit der Auswertung diese Tabelle ist der direkte Vergleich der durchschnittlichen Precision-Ergebnisse (bei mehreren Anfragen pro Kategorie werden Mittelwerte gebildet) über alle Läufe (vgl. Tabelle 10.4)

Diese Form der Auswertung führt jedoch zu keinem klaren Bild oder einer nachvollziehbaren Tendenz. Es lässt sich allenfalls feststellen, dass die Anfragen mit überregionalen geopolitischen Angaben und vagen Themenangaben über alle Evaluierungsläufe hinweg sehr schlecht, während die Anfragen mit Nennungen von oberbegriffsartigen Themenangaben und Staatenbünden über alle Läufe sehr gut abgeschnitten haben. Die Kombination von oberbegriffsartigen Themenangaben mit überregionalen geopolitischen Angaben und die Kombination von präzisen Themenangaben mit der Nennung von Staatenbünden ist dagegen sehr erfolgreich bei den Evm-Läufen.

10.1.3. Mehrsprachige Retrievalleistung

Im Rahmen der Untersuchung der mehrsprachigen Retrievalleistung zeigt sich deutlich, dass durch die verstärkte Nutzung des kontrollierten Vokabulars und des Einsatzes der Übersetzung die mehrsprachige Retrievalleistung drastisch gesteigert werden konnte. Nicht nur, dass die relevanten fremdsprachlichen Dokumente insgesamt von 202 (SwpBase1) bzw. 330 (SwpBase2) auf zwischen 510 (SwpEvm3) und 581 (SwpEvm1) um durchschnittlich (Veränderung durchschnittlicher fremdsprachlicher Recall SwpBase-

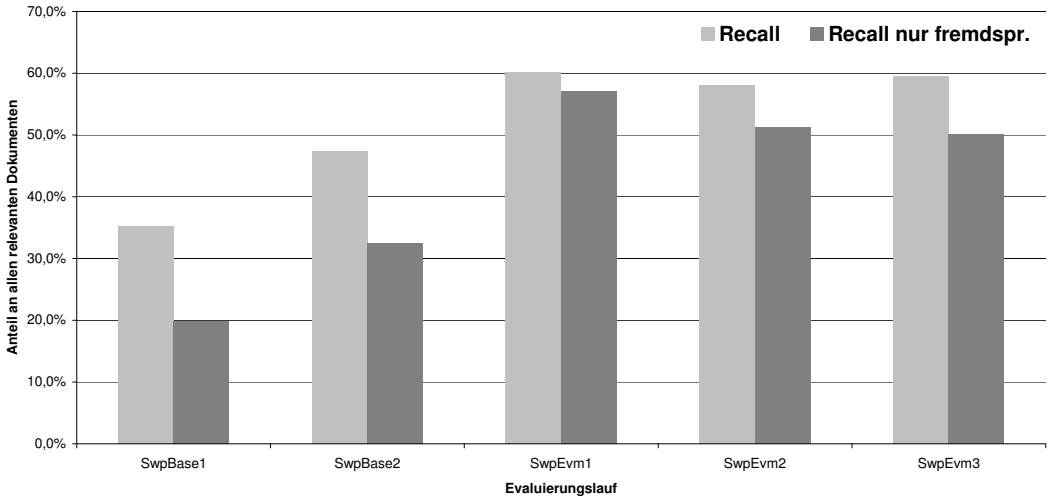

Abbildung 10.2.: Vergleich der mehrsprachigen Retrievalleistung nach Recall

	SwpBase1	SwpBase2	SwpEvm1	SwpEvm2	SwpEvm3
Retrieved	5000	5000	5000	5000	5000
Relevant	2039	2039	2039	2039	2039
Relevant fremdsprachlich	1018	1018	1018	1018	1018
Rel-Ret	717	965	1226	1182	1211
Rel-Ret fremdsprachlich	202	330	580	521	510

Tabelle 10.5.: Recall-Ergebnisse über alle Anfragen und alle Evaluierungsläufe

Gruppe zu durchschnittlichem fremdsprachlichen Recall SwpEvm-Gruppe) 102% gestiegen sind: Insgesamt konnte auf diesem Wege auch der ursprüngliche Anteil der durch die SwpBase-Gruppe gefundenen, relevanten fremdsprachlichen Dokumente an allen, durch diese Gruppe gefundenen, relevanten Dokumente von 31,2% auf in den SwpEvm-Läufen erreichten 44,5% gesteigert werden.

Deutlich zeigt sich ebenfalls die Steigerung des Anteils der je Evaluierungslauf gefundenen relevanten, fremdsprachlichen Dokumente an der Anzahl aller solcher fremdsprachlichen Dokumente. Hierbei konnten die SwpEvm-Läufe gegenüber den SwpBase-Läufen noch deutlicher zulegen, als im gesamten Vergleich mit allen Sprachen: Der durchschnittliche Anteil über die SwpBase-Gruppe liegt bei 26,1%, während der Anteil der SwpEvm-Gruppe bei 52,8% liegt. Die SwpEvm-Gruppe konnte entsprechend durch den Einsatz der zusätzlichen Module zusätzlich einen deutlich größeren Anteil von allen relevanten, fremdsprachlichen Dokumenten auffinden.

In dieser Auswertung zeigt sich ebenfalls erneut der starke Unterschied zwischen der Leistungsfähigkeit einer Anfrage gänzlich ohne Metadaten, wie im Lauf SwpBase1 simuliert, gegenüber einer Anfrage mit einer grundlegenden Einbindung von Metadaten

in SwpBase2: Der Anteil relevanter, fremdsprachlicher Dokumente an allen relevanten Dokumenten ist dadurch von 28,2% auf 34,2% gestiegen.

Darüber hinaus bestätigt sich ebenfalls die Leistungsfähigkeit von SwpEvm1, der sich erneut als insgesamt stärkster Evaluierungslauf herausstellt. Gegenüber dem zweitplatzierten SwpEvm2 erreicht SwpEvm1 im Recall von fremdsprachlichen Dokumenten ein um 11,3% besseres Ergebnis.

10.1.4. Leistungsfähigkeit der einzelnen Module

Um nicht nur die Leistungsfähigkeit des gesamten IR-Systems auszuwerten, sondern auch die Anteile der einzelnen Module an der Retrievalqualität annäherungsweise zu bewerten, soll in diesem Abschnitt eine Auswertung der Evaluierung beschrieben werden, die diese Anteile misst.

Basierend auf dem erfolgreichsten SwpEvm-Lauf, SwpEvm1, wurden in drei zusätzlichen Läufen ausschließlich die Passagen der jeweilig untersuchten Module zusätzlich zur grundlegenden Anfrage, die auf die Feldern *title* und *abstract* gerichtet wurde (entsprechend SwpBase1), angewendet.

Diese Auswertung kann nur annähernde Ergebnisse liefern, da ggf. nicht alle gefundenen Dokumente in der Relevanzbewertung ausgewertet wurden. Da Dokumente, die innerhalb dieser Vergleichsläufe gefunden werden, aber noch nicht vorher im Rahmen der Relevanzbewertung ausgewertet wurden, als irrelevant gelten, ist es möglich, dass, je nachdem, wie hoch der Anteil an nicht relevanzbewerteten Dokumenten ist, die tatsächlichen Ergebnisse der Vergleichsläufe von BRF, EVM und Translator besser abschneiden. Dennoch gibt die Auswertung einen Anhaltspunkt darüber, welches Modul hauptsächlich die Steigerung der Retrievalleistung erbracht hat.

Da die drei Modulanfragen SwpEvm_only, SwpBrf_only und SwpTranslator_only dem einfacheren SwpBase1 zuzüglich von Anfragepassagen aus den einzelnen Modulen entsprechen, wird dieser Lauf auch als Maßstab für die Verbesserung der Retrievalqualität herangezogen. Als Referenz in diesem Vergleich gilt der beste Evaluierungslauf, SwpEvm1, da in ihm alle drei Module zusammen die Anfrage erweitern.

Während die Ergebnisse natürlich nicht exakt kumulativ zu verstehen sind, gibt diese Übersicht jedoch einen gewissen Überblick darüber, welche Module welchen Anteil an der Verbesserung der Retrievalleistung hatten: Es zeigt sich, dass SwpEvm_only nahezu die Leistungsfähigkeit von SwpEvm1 in der Precision (Differenz von 0,015) und im Recall (Differenz von 78 Dokumenten oder 6,8%) erreicht. SwpBrf_only kann sich dagegen zwar recht deutlich von SwpBase1 abheben, erreicht aber keine umfassende Steigerung und ist in der Leistung nicht vergleichbar mit SwpEvm1. SwpTranslator_only erreicht ein schlechteres Recall-Ergebnis als SwpBase1 und auch nur eine

	SwpBase1	SwpTranslator_only	SwpBrf_only	SwpEvm_only	SwpEvm1
Retrieved	5000	5000	5000	5000	5000
Relevant	2039	2039	2039	2039	2039
Rel-Ret	717	694	828	1148	1226
0	0,5815	0,5873	0,6078	0,5717	0,6787
0,1	0,2596	0,2934	0,3231	0,3872	0,4085
0,2	0,1945	0,2289	0,2561	0,3567	0,3622
0,3	0,1541	0,1659	0,2062	0,3193	0,3252
0,4	0,1106	0,126	0,1675	0,2647	0,2932
0,5	0,0671	0,0631	0,1032	0,2199	0,2349
0,6	0,0413	0,0522	0,0218	0,1557	0,1784
0,7	0,0057	0,0078	0,0137	0,0974	0,1213
0,8	0	0	0,0115	0,0119	0,033
0,9	0	0	0	0	0,0123
1	0	0	0	0	0,0017
Durchschnitt	0,0985	0,1114	0,1291	0,1984	0,2138

Tabelle 10.6.: Precision-Ergebnisse über die verschiedenen Module

minimale Steigerung der Precision von 0,098 auf 0,111. Es zeigt sich jedoch, dass die Precision unter den ersten 10% des Recalls nur dann deutlich über die Leistung von SwpBase1 steigt, wenn alle drei Module in SwpEvm1 zusammenarbeiten.

Diese Steigerung hängt also offensichtlich mit der Nutzung der Metadaten zusammen, die durch das EVM der Suchanfrage ergänzt werden. Allerdings fällt ebenfalls auf, dass im direkten Vergleich zu SwpBase2 der Evaluierungslauf SwpEvm_only sich nicht deutlich in der Precision (0,0232) absetzen konnte. Allerdings konnte die Recall-Leistung durch den exklusiven Einsatz des EVM um rund 20% von 965 auf 1148 gesteigert werden. Diese Leistungsveränderung geht also konform mit dem bereits erwarteten Verhalten, dass sich durch die Ergänzung einer größeren Anzahl von mehrheitlich geeigneten Termen der Recall stärker steigt als die Precision.

Insgesamt ist davon auszugehen, dass die unveränderte Gewichtung der Originalanfrage bei einer deutlich unterschiedlichen Anzahl von ergänzten Begriffen die Ergebnisse der einzelnen Anfragen beeinflusste. Ungeachtet dessen lässt sich durch diese Ergebnisse belegen, dass das EVM durch seine Anfragenergänzung im Rahmen der gesamten Evaluierung den effektivsten Beitrag geleistet hat: Die Leistung des Retrievalsystems hat sich im Rahmen von SwpEvm3 nicht deutlich verändert, obwohl die Anzahl der hinzugefügten Terme etwa halbiert wurde.

Während also EVM und BRF Verbesserungen erzielen konnten, war Leistungssteigerung durch das Übersetzungsmodul nicht deutlich messbar. Für den Translator ergibt eine Auswertung der Protokolle des Suchprozess ebenfalls ein gemischtes Bild: Während einige Anfragen (besonders die konkreten Fragen mit Staatennamen und klaren, thematischen Schlagwörtern) recht zuverlässig durch den Translator übersetzt werden konnten, wurden vage oder oberbegriffsartige Begriffe wie „Sicherheit“ meist um

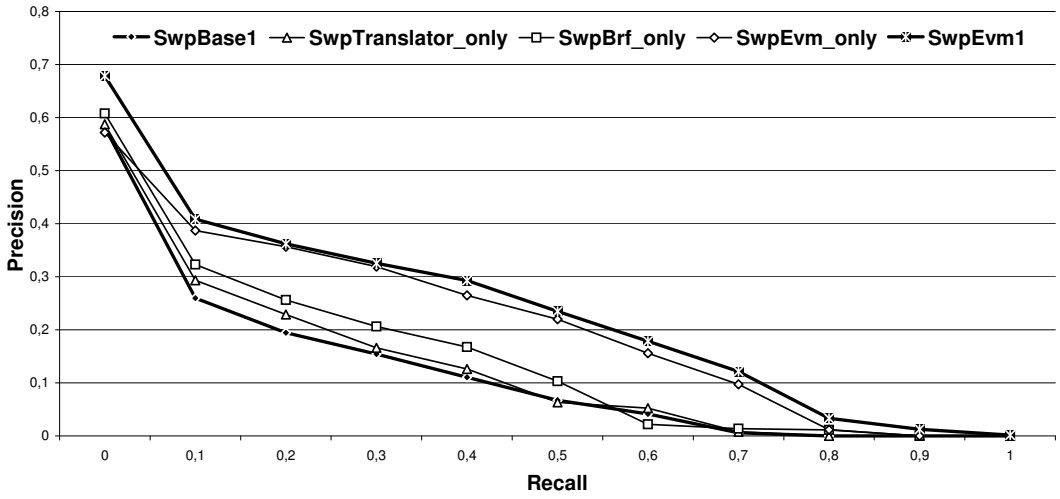

Abbildung 10.3.: Vergleich der einzelnen Module

Übersetzungen ergänzt, die diesen Begriff als Teil einer Phrase enthalten (zum Beispiel „Technische Sicherheit“, „Wirtschaftliche Sicherheit, etc.). Dies wurde dadurch verursacht, dass im Übersetzungsprozess nach dem Term im Thesaurus gesucht wird und nur die Übersetzung der Begriffe, die von Lucene mit 1,0 bewertet wurden, zurückgegeben wurden. Dieser maximale Wert konnte aber in dieser Konstellation offensichtlich auch durch Phrasen erreicht werden, bei dem nur ein Wort das Gesuchte darstellte. Während dadurch der Anfrage auch nützliche Phrasen ergänzt wurden („sicherheitspolitische“ wurde beispielsweise mit Übersetzung des Begriffs „Sicherheitspolitische Faktoren“ erweitert), war dies grundsätzlich nicht das Ziel der Übersetzung. Ein weiterer Faktor der Einschränkung der Leistungsfähigkeit des Translators war darüber hinaus, dass die hinzugefügten Terme nur auf das Feld *abstract* und nicht auf beide Felder *abstract* und *title* gerichtet wurden. Da viele fremdsprachige Dokumente auch fremdsprachliche Titel haben, hätte dies die Leistung weiter steigern können.

10.2. Evaluierung für GIRT3

10.2.1. German Indexing and Retrieval Database 3

Das vorgestellte System wurde nicht nur auf der Datenbasis des Fachinformationsverbundes angewendet, sondern auch auf die German Indexing and Retrieval Testdatabase (GIRT). Auch bei GIRT handelt es sich um eine umfassend durch Metadaten erschlossene Referenzdatenbank (vgl. Tabelle 10.7), entsprechend lassen sich

ebenfalls Deskriptoren zur Erweiterung der Suchanfrage verwenden. Die in der GIRT-Datenbank verzeichneten Dokumente beziehen sich inhaltlich auf soziologische Themen wie Industrie- und Betriebssoziologie oder Migration und ethnische Minderheiten.

	Anzahl	Anteil an gesamt
Dokumente insgesamt	76.128	100,0%
Dokumente mit Titel	76.128	100,0%
Dokumente mit Klassifikationsangaben	76.128	100,0%
Dokumente mit thematischen Deskriptoren	76.126	99,9%
Dokumente mit Zusammenfassungen	73.291	96,3%
	Anzahl gesamt	Anzahl pro erschl. Dokument
Verzeichnete Deskriptoren	755.333	9,9
Verzeichnete Klassifikationsangaben	169.064	2,2

Tabelle 10.7.: Erschließung des GIRT-Datenbestands durch Metadaten

Alle fünf Evaluierungsläufe, die in Abschnitt 9.3.2 ab Seite 61 vorgestellt wurden, werden auf den Datensatz GIRT3 angewendet. Während GIRT4 ebenfalls vorliegt, fiel die Entscheidung für die parallele Evaluierung zugunsten des älteren GIRT3-Datensatzes, da sie eine größere systematische Ähnlichkeit zur Datenbasis des Fachinformationsverbundes hat: GIRT4 umfasst zwar mit 151.000 unterschiedlichen Dokumenten, die sowohl in englisch als auch in deutsch vorliegen (entsprechend 302.000 Datensätze), deutlich mehr Datenmaterial, allerdings wurden bei der Übersetzung der Dokumente ins Englische auch die Deskriptoren und die Bezeichnungen der Klassifikation übersetzt. Eine entsprechende Anpassung des System wäre zwar möglich gewesen, hätte die Ergebnisse aber durch die Implementierung eines zusätzlichen Moduls nicht im gleichen Maße vergleichbar gemacht.

10.2.2. Anpassung des Systems

Das System wurde im Rahmen der Evaluierung auf GIRT3 in mehreren Faktoren angepasst. Zunächst wurde das Parsing und die Indexierung der rund 156 MB großen Rohdatendatei auf die XML-Architektur angepasst, entsprechend wird nun auf die Felder *docno* (Dokumentennummer), *title* (Dokumententitel), *text* (Zusammenfassung), *controlled-term* (Deskriptoren des kontrollierten Vokabulars) und *classification* (Gruppenbezeichnungen der GIRT-Klassifikation) indexiert. Im Rahmen der Indexierung wurde ebenfalls die Stoppwortliste überarbeitet und an die besonders häufig auftretenden Terme angepasst. Ergänzend wurden einige Stoppworte aus den Evaluierungsanfragen übernommen.

Da es in dieser Datenbasis wegen des thematischen Fokus auf soziologische Themen keine Metainformationen zu einem eventuellen geopolitischen Bezug des betreffenden Dokuments gibt, wurden im Prozess des EVM alle Kaskadenelemente deaktiviert, die

	GIRT_SwpBase1	GIRT_SwpBase2	GIRT_SwpEvm1	GIRT_SwpEvm2	GIRT_SwpEvm3
Retrieved	5000	5000	5000	5000	5000
Relevant	1111	1111	1111	1111	1111
Rel-Ret	416	517	627	599	562
0	0,8063	0,8131	0,8426	0,8574	0,7915
0,1	0,6112	0,6369	0,7005	0,6846	0,6386
0,2	0,3472	0,4166	0,5695	0,5184	0,4952
0,3	0,2156	0,3452	0,4323	0,3883	0,3357
0,4	0,1403	0,2357	0,3178	0,2842	0,2423
0,5	0,0714	0,1360	0,2433	0,2199	0,1563
0,6	0,0426	0,0896	0,1425	0,0977	0,1078
0,7	0,0426	0,0624	0,0593	0,0480	0,0540
0,8	0,0398	0,0335	0,0240	0,0027	0,0109
0,9	0,0038	0,0091	0,0025	0,0000	0,0000
1	0	0	0	0	0
Durchschnitt	0,1879	0,2309	0,2843	0,2575	0,2350

Tabelle 10.8.: Retrievalleistung über alle Evaluierungsläufe für GIRT

diesen Typ von Metainformation betreffen. Dadurch wurde die Kaskade im Vergleich zum Original, das auf den FIV-Bestand angewendet wurde, um vier Elemente verkürzt.

Die verwendeten Evaluierungsanfragen sind Anfragen 26-51 der GIRT- Anfragesammlung. Es gibt im Rahmen dieser Evaluierung umfassende Anfragebeschreibungen (die Anfragen bestehen aus einem Titel, einer kurzen und einer detaillierteren Beschreibung). Um eine möglichst große Anzahl von relevanten Termen in die Anfragen zu übernehmen, wurden alle drei Teile der Anfrage eingesetzt.

10.2.3. Ergebnisse der Evaluierung

Die Ergebnisse der GIRT-Evaluierung zeigen eine deutliche Steigerung der Retrievalleistung des vorgestellten Systems gegenüber der FIV-Evaluierung. Sowohl die Gruppe der SwpBase-Läufe erreichen deutlich bessere Retrievalergebnisse, als auch die Gruppe der SwpEvm-Läufe.

GIRT_SwpBase1 vs. GIRT_SwpBase2

Die Retrievalqualität sich steigert von GIRT_SwpBase1 zu GIRT_SwpBase2 sowohl im Recall (37,4% auf 46,5%) als auch in der Precision (18,8% auf 23,1%). Damit ist die Steigerung zwischen dem Lauf ohne Anwendung von Metadatenfeldern und dem Lauf mit Nutzung von Metadatenfeldern weniger deutlich gestiegen als bei der Anwendung auf den FIV-Datensatz.

Es ist zu vermuten, dass dies damit zusammenhängt, dass beinahe alle Dokumente in GIRT über eine Zusammenfassung verfügen und damit der Anteil der Dokumente

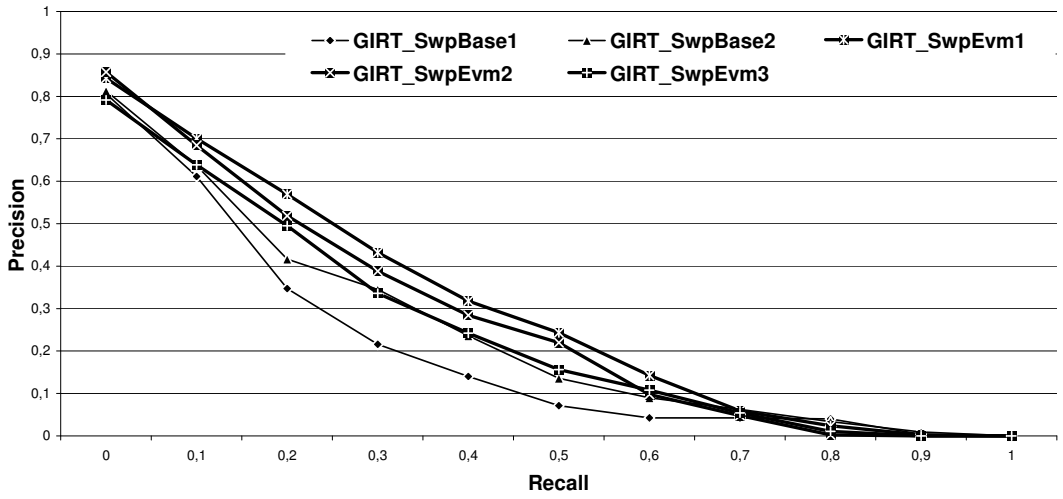

Abbildung 10.4.: Vergleich Evaluierungsläufe auf GIRT

mit Zusammenfassung rund viermal so hoch ist wie bei FIV. Entsprechend ist die Anwendung einer Suchanfrage auf die Felder *title* und *text* deutlich effektiver geworden.

GIRT_SwpEvm1 vs. GIRT_SwpEvm2 vs. GIRT_SwpEvm3

Erneut zeigt sich SwpEvm1, diesmal in abgewandelter Form als GIRT_SwpEvm1, als hochwertigster Ansatz (56,4% Recall, 28,4% Precision) zur Anfrageergänzung. Im Rahmen der GIRT-Evaluierung zeigt sich allerdings ein stärkerer SwpEvm2-Lauf (53,9% Recall, 25,8% Precision). SwpEvm3 ist sowohl in Recall (50,6%) als auch in Precision (23,5%) der schwächste Lauf dieser Gruppe.

Im Gegensatz zur Evaluierung für den FIV-Bestand ergibt sich für GIRT also eine, mit 4,9% in der Precision und 5,8% im Recall, deutlichere Bandbreite der Leistungsfähigkeit unter den SwpEvm-Läufen.

GIRT_SwpBase vs. GIRT_SwpEvm

Im Vergleich zur Evaluierung auf den FIV-Datenbestand tritt die Gruppe der SwpBase-Läufe in Anwendung auf GIRT3 homogener auf, dies erleichtert den Vergleich der Gruppen.

Die Recall-Leistung der Gruppe der SwpBase-Läufe liegt mit 416 und 517 von 1111 relevanten Dokumenten deutlich (im Mittel um 27,8%) unter der Recall-Leistung der SwpEvm-Läufe (562, 599 und 627 Dokumente).

Die Precision konnte von durchschnittlich 20,9% auf 25,9% durch die Anwendung der zusätzlichen Module gesteigert werden.

Zusammenfassung und Schlussfolgerungen aus den Ergebnissen

GIRT_SwpBase1 hat deutlich besser als SwpBase1 für den FIV-Datenbestand abgeschnitten. Dies ist offensichtlich auf die nahezu vollständige Erschließung von GIRT3 durch Zusammenfassungen zurückzuführen. Da dieser Evaluierungslauf nur die Felder Titel und Text anspricht, steigt durch die Verfügbarkeit von Zusammenfassungen im Feld Text eindeutig die Leistungsfähigkeit.

SwpBase2 konnte die Leistungsfähigkeit durch die zusätzliche Anwendung der Anfragen auf die kontrollierten Felder der Deskriptoren und der Klassifikation deutlich steigern.

Sowohl GIRT_SwpEvm1 als auch GIRT_SwpEvm2 konnten sich deutlich von der Gruppe der GIRT_SwpBase-Läufe absetzen, während GIRT_SwpEvm3 besonders in der Precision nicht deutlich über die Leistung von GIRT_SwpBase2 hinauskommt. Dies mag auf die Kombination der hohen Anzahl der in der statistischen Auswertung berücksichtigten Dokumente und der gleichmäßigen, einfachen Gewichtung der EVM-Terme zurückzuführen sein (vgl. Abschnitt 9.3.2 auf Seite 61). In dieser Evaluierung zeigten sich eindeutig die Läufe mit gewichteten EVM-Termen im Vorteil. Darüber hinaus scheint eine größere Anzahl von ergänzten Deskriptoren positive Auswirkungen auf die Precision zu haben. Im Unterschied zwischen GIRT_SwpEvm1 und GIRT_SwpEvm2 zeigt sich, wie bereits in der Evaluierung auf den FIV-Bestand, dass eine geringere Anzahl von berücksichtigten Dokumenten vielversprechender ist.

10.3. Zusammenfassung der Ergebnisse

10.3.1. Einzelne Ergebnisse

Um in gebotener Kürze die wichtigsten Ergebnisse der umfassenden Evaluierung des Systems zu rekapitulieren, werden im Folgenden stichpunktartig die wichtigsten Aspekte der Ergebnisse zusammengefasst:

Ergebnisse FIV gesamt

- Die Precision konnte generell durch den Einsatz der zusätzlichen Module von 0,985 (SwpBase1) bzw. 0,1752 (SwpBase2) auf 0,2138 (SwpEvm) gesteigert werden. Der Recall stieg von 33% (SwpBase1) bzw. 47% (SwpBase2) auf 60% (SwpBase1).
- SwpBase2 ist durch die Anwendung der Anfragenterme auf die Metadatenfelder deutlich leistungsfähiger geworden als SwpBase1 und ist daher eher mit den Evm-Läufen zu vergleichen.
- Die drei Evm-Läufe unterscheiden sich nicht deutlich in ihrer Leistungsfähigkeit in dieser ganzheitlichen Analyse.

Ergebnisse FIV anfragenorientiert

- In den Evm-Läufen konnte die Anzahl der nach Precision überdurchschnittlich abschneidenden Anfragen gesteigert werden. SwpEvm3 erreicht hierbei das beste Ergebnis mit 44% der Anfragen mit überdurchschnittlichen Werten (gegenüber SwpBase mit nur 28%). Nach einer Gegenprobe der Anfragenergebnisse im Lauf SwpEvm_only konnte bestätigt werden, dass dieses Verhalten hauptsächlich vom EVM erzielt wurde, die beiden anderen Module hatten hierauf wenig Einfluss: SwpBrf_only erreichte genau das gleiche Verhältnis von überdurchschnittlich abschließenden Anfragen wie die SwpBase-Läufe.
- Die Anzahl der Anfragen mit einem schlechteren Ergebnis als einem Drittel des Durchschnitts sinkt von 36% bei SwpBase1 auf 20% bei SwpEvm3.
- SwpEvm1 erreicht mit 32% den größten Anteil an hervorragend (150% des Durschnitts) abschneidenden Anfragen.
- Anfragen 16 und 22 schneiden über alle Läufe mit auffälligem Abstand am besten ab. Es ist zu vermuten, dass dies mit der thematischen Ausrichtung der Frage und der offenen Formulierung zusammenhängt.

- Anfragen mit einem überdurchschnittlichen Ergebnis bei SwpBase1 erreichen bei den Evm-Läufen nur unterdurchschnittliche Ergebnisse und umgekehrt.

- Anfragen mit einem überdurchschnittlichen Ergebnis bei SwpBase1 konnten zwei Kategorien des Anfragenschemas zugeordnet werden (keine geopolitische Angaben & präzise Themenangaben bzw. oberbegriffsartige Themenangaben).

- Anfragen, die in den Evm-Läufen besonders gute Ergebnisse erzielten, hatten überwiegend vage Themenangaben. Anfragen, die in den Base-Läufen besonders gute Ergebnisse erzielten, hatten entweder präzise oder oberbegriffsartige Themenangaben.

- Sehr vage thematische und geopolitische Angaben erzielen schlechte Ergebnisse über alle Evaluierungsläufe.

Ergebnisse FIV Mehrsprachigkeit

- Die Anzahl an relevanten, fremdsprachigen Dokumenten konnte bereits durch SwpBase2 gegenüber SwpBase1 drastisch gesteigert werden.

- In den Evm-Läufen nahm der Recall weiter zu und ist in der Gruppe der fremdsprachlichen Dokumenten (Anteil Rel-Ret fremdsprachlich an Rel fremdsprachlich) mit 60% nahezu genauso hoch wie der Recall der deutschen Dokumente.

Ergebnisse FIV Moduluntersuchung

- Die Leistung einer ausschließlich durch das EVM-Modul erweiterten Anfrage kommt nahe an die Leistung einer Anfrage mit allen Modulpassagen heran (Differenz von nur 0,015). Die Leistung des EVM ist also maßgeblich für die Retrievalqualität in den SwpEvm-Evaluierungsläufen.

- Unter den ersten 10% des Recalls steigt die Precision nur bei Kombination aller Module um weitere rund 10%.

- Der Translator hat den geringsten Anteil an der Steigerung der Retrievalleistung (Differenz von rund 0,01 zu SwpBase1).

- Das Blind Relevance Feedback für den Freitext steigert die Anfragequalität in der Precision um rund 0,03.

Ergebnisse GIRT gesamt

- Alle fünf Läufe schneiden bei GIRT besser ab als bei FIV.
- Die Differenz der durchschnittlichen Precision zwischen SwpBase1 bzw. SwpBase2 zum besten Evm-Lauf SwpEvm1 liegt bei 0,0964 bzw. 0,0534, damit kann die Verbesserung als signifikant bezeichnet werden.
- Es wird angenommen, dass der größere Anteil an Dokumenten mit Zusammenfassungen eine Verbesserung EVM-Leistung bewirkt hat. Ein weiterer, wichtiger Faktor war darüber hinaus die Verkürzung der Kaskade auf 5 Elemente, die keine Wiederverwendung der bereits extrahierten Deskriptoren vorsieht.
- Der Recall erreicht im Bestfall (SwpEvm1) maximal 56,4%, was darauf hindeutet, dass viele unterschiedliche, relevante Dokumente in den diversen Läufen gefunden wurden.

10.3.2. Erkenntnisse bezüglich der zentralen Fragestellungen

In Abschnitt I auf Seite xv wurden sechs spezifische Fragestellungen formuliert, die nach einem ersten Überblick über die Datengrundlage und vor der Entwicklung des Systems als zentrale Richtlinien für das Projekt fungierten.

1. Welche Ansätze für den Information Retrieval Prozess lassen sich aus dem speziellen Anwendungsfall des Retrievals von Dokumenten mit inhaltlichem Bezug auf Außen- und Sicherheitspolitik ableiten?

Für das Textretrieval in der spezifischen Domain der Texte mit außen- und sicherheitspolitischem Bezug lassen sich mehrere wichtige Punkte festhalten: Sowohl Anfragen, als auch die verzeichneten Texte, beinhalten in einem Großteil der Fälle sowohl einen thematischen als auch einen geopolitischen Bezug (vgl. Analyse in VI auf Seite 110 und die dazugehörige Auswertung 9.2 ab Seite 56). Diese Kombination von Bezugssystemen ergibt sich zwangsläufig aus der Domäne, auf die sich die Anfragen und Dokumente beziehen. Sowohl die Orientierung auf einen einzelnen Staat oder internationalen Bezug zwischen zwei oder mehreren Staaten, Staatenbünde oder Regionen, als auch einen thematischen Bezug, also den Aspekt, unter der die uni-, bi- oder multilateralen Beziehungen zum Tragen kommen, werden betrachtet und so beide Bezugssysteme verknüpft. Beide Bezugssysteme sollten bei der Entwicklung eines Information Retrieval Systems in dieser Domäne besondere Beachtung finden. Zusätzlich zu den

vorherrschenden, thematischen und geopolitischen Bezugssystem werden in 20% der Anfragen explizite temporale Angaben gemacht. Darüber hinaus sind temporale Angaben in einem Großteil der Anfragen meist implizit vorhanden (siehe beispielsweise Anfrage 18). Dies ist dem zeitgeschichtlichen Aspekt der Politikwissenschaft geschuldet: Die Politikwissenschaft befasst sich, ähnlich der Geschichtswissenschaften, mit Ereignissen, Prozessen, Entwicklungen und Zuständen die immer im Kontext der Zeit, in der sie sich ereignen, zu sehen sind. Für die thematischen und geopolitischen Bezugssysteme ergibt sich eine gewisse Bandbreite des Vokabulars, welches von der konkreten und klar definierten („Nigeria“, „Atomkonflikt“) bis hin zur uneindeutigen und vagen Semantik („Probleme“, „mediterraner Raum“) reicht. In diesem Rahmen können durch ein EVM aktivierte Klassifikationen und Thesauri helfen, vage und schlecht einsetzbare Oberbegriffe in konkretere Angaben zu wandeln. Im Gegenzug scheint eine Erweiterung von Anfragen, die ohnehin geeignetes Vokabular mit sich bringen, weniger erfolgreich.

2. Welche Retrievalstrategien lassen sich für die Entwicklung eines Retrievalsystems für Referenzdatenbanken ableiten?

Referenzdatenbanken sind im Gegensatz zu Volltextdatenbanken üblicherweise besser durch Metadaten erschlossen. Die Datenbasis des Fachinformationsverbundes ist ein gutes Beispiel für eine hervorragende Erschließung durch kontrolliertes Vokabular in Form von Deskriptoren und einer Klassifikation (vgl. Analyse der Datengrundlage in 4.1.4 ab Seite 30). Diese Metadaten lassen sich wie in Abschnitt 10.1.1 belegt durch ein dynamisches Entry Vocabulary Modul (vgl. Abschnitt 5) im Rahmen des Retrievalprozesses nutzen. Der eindeutigste Beleg für die Steigerung der Retrievalleistung ist die Leistungssteigerung von Evaluierungslauf SwpBase1 ohne jegliche Nutzung der Metadaten auf SwpBase2 mit einfacher Nutzung der Metadaten. Der positive Einfluss der Metadaten auf das Retrievalergebnis konnte durch das EVM weiter gesteigert werden.

3. Lässt sich die Retrievalleistung eines IR-Systems durch die Anwendung eines Entry Vocabulary Moduls verbessern?

In Abschnitt 10.1.1 und detaillierter in Abschnitt 10.1.4 ab Seite 82 wurde die Leistungssteigerung der Evaluierungsläufe mit EVM gegenüber den Läufen ohne zusätzliche Module belegt.

Die Auswertung in 10.1.4 auf Seite 82 belegt, dass die Leistungssteigerung in dieser Evaluierung zum Großteil vom EVM getragen wurde.

4. Wie verändert sich die Retrievalleistung eines IR-Systems mit Entry Vocabulary Modul, wenn die Parameter des Systems geändert werden?

Nach einer ersten Analyse der einzelnen Evaluierungsläufe zeigte sich, dass die Veränderung der gesamten Retrievalleistung durch das Anpassen der Parameter des EVM sich in engen Grenzen hielt: Während sich die Precision der drei EVM-Läufe zwischen SwpEvm1 und SwpEvm2 maximal um 1,6% unterscheidet, liegen auch die Ergebnisse im Recall mit einer maximalen Differenz zwischen SwpEvm1 und SwpEvm2 von 3,7% in einem Rahmen, der keine eindeutigen Aussagen zulässt. Danach ließe sich feststellen, dass SwpEvm1 mit den Parametern 0,3 für den *cutOffScore* und 30 für *consideredDocs* der erfolgreichste Lauf war, also eine große Anzahl von hinzugefügten Deskriptoren besonders erfolgversprechend ist.

Allerdings zeigt sich in einer tiefergehenden, anfrageorientierten Analyse in Abschnitt 10.1.2 ab Seite 73, dass bei SwpEvm1 nur 32% der Anfragen überdurchschnittliche Ergebnisse erzielen konnten. Diese Anfragen hatten alle eine um 50% höhere Precision als der Durchschnittswert. Dagegen zeigte sich SwpEvm3 (0,6/100) als ein homogenerer Evaluierungslauf: 44% der Anfragen konnten überdurchschnittlich abschneiden. Außerdem konnten der Anteil der Anfragen, die nur unter einem Drittel der Durchschnittsleistung lagen, von 24% auf 20% gesenkt werden. Das Hinzufügen von vielen Deskriptoren aus einer geringeren Anzahl von untersuchten Dokumenten führte also zu einem tendenziell besseren Gesamtergebnis, hatte aber als Nebeneffekt, dass nur wenige Anfragen deutlich profitierten. Das Hinzufügen von wenigen Deskriptoren aus vielen untersuchten Dokumenten führt dagegen zu einem umgekehrten Effekt.

SwpEvm2 konnte dagegen mit den Parametern 0,3/100 nicht überzeugen. Die Kombination von einer großen Anzahl von untersuchten Dokumenten und einer großen Anzahl von übergebenen Deskriptoren durch den geringeren Schwellenwert führte dazu, dass ebenfalls mehr unnütze Deskriptoren der Anfrage hinzugefügt wurden. Zwar stieg der überdurchschnittlich abschneidende Anteil der Anfragen gegenüber SwpEvm1 deutlich, allerdings stieg im ähnlichen Maße der Anteil an Anfragen, die nur weniger als ein Drittel der durchschnittlichen Retrievalleistung erreichen konnten.

Die Analyse der untersuchten Parameter führt zu der Vermutung, dass die Kombination von einer geringeren Anzahl von untersuchten Dokumenten und einer geringeren Anzahl von effektiv der Anfrage hinzugefügten Deskriptoren möglicherweise zu einer Kombination von homogener Retrievalleistung und höherem Gesamtergebnis führt. Dies wäre in einer weiteren Untersuchung zu beachten.

5. Welchen Einfluss haben die eingesetzten Module auf die mehrsprachige Retrievalleistung?

Die eingesetzten Module konnten, wie in Abschnitt 10.1.3 auf Seite 80 dokumentiert, die mehrsprachige Retrievalleistung deutlich steigern. Aufgrund der eingesetzten Module wurde der Anteil von relevanten, fremdsprachlichen Dokumenten verdoppelt. Damit ist eine eindeutige Verbesserung der Retrievalleistung über Sprachgrenzen hinweg mit Hilfe der zusätzlich eingesetzten Module, vor allem des Translators und des EVM, realisiert worden.

6. Welche zusätzlichen Ansätze zur Steigerung der Retrievalleistung sind während der Entwicklung des Systems deutlich geworden?

Im Abschnitt 11 ab Seite 96 werden diverse Verbesserungsvorschläge und Ansätze zur Ergänzung des vorgestellten Systems genannt.

Teil V.

Ausblick und Fazit

11. Ansätze für weitere Verbesserungen

Während in dieser Arbeit belegt werden konnte, dass das Entry Vocabulary Modul die Leistungsfähigkeit des Retrievalsystems steigern konnte, ist das vorliegende System mit den ebenfalls vorliegenden Evaluierungsanfragen noch nicht in dem Maße leistungsfähig, dass so natürlichsprachlich verfasste Anfragen verwendet werden können, um höchste Retrievalleistungen zu erreichen. Da es sich um ein prototypisches System mit dem Fokus auf der Evaluierung eines einzelnen Moduls handelt, ist dies grundsätzlich zwar kein Misserfolg, ein in seiner Retrievalleistung konkurrenzfähiges System soll aber die Zielsetzung für die folgenden Verbesserungsvorschläge und zu erörternden Optionen sein.

Es gibt vielerlei Ansätze, mit denen sich die Leistungsfähigkeit des vorliegenden Retrievalsystems verbessern lässt. Einige Ansätze umfassen die Verbesserung von bestehenden Methoden, während sich andere Ansätze darauf konzentrieren, weitere Verfahren zu ergänzen.

11.1. Analyse der Schwachstellen und Vorschläge für Verbesserungen

Das vorgestellte System hatte an mehreren Punkten Schwachstellen, die im Sinne der weiteren Verbesserung der Retrievalqualität behoben werden müssen. Es ist zu erwarten, dass sich durch die Überarbeitung des Systems die Retrievalleistung signifikant steigern lässt.

11.1.1. Verbesserung der Anfrageformulierung des EVM

Das EVM gibt im vorliegenden System Anfragepassagen entsprechend dem Beispiel `(Vereinigte Staaten)^0.7` zurück. Da es sich bei diesem und allen Deskriptoren um exakte Phrasen handelt, sollte die Anfrage korrekterweise „`Vereinigte Staaten`“`^0.7`

lauten, da sonst auch Dokumente mit einem Deskriptor, der das Wort „Vereinigte“ beinhaltet, zu einem gewissen Maß als relevant betrachtet werden. Die Deskriptoren wurden also nicht phrasenexakt eingesetzt. Es ist zu erwarten, dass eine dahingehende Verbesserung des Systems die Retrievalleistung weiter steigern kann.

11.1.2. Einsatz eines geeigneteren Stemmers

Im Rahmen der Entwicklung des Information Retrieval Systems kam der Lucene StandardAnalyzer zum Einsatz. Dieser Analyzer ist grundsätzlich gut geeignet, um eine Grundformenreduktion für westliche Sprachen umzusetzen, er ist aber nicht speziell für die deutsche Sprache optimiert. Der Einsatz des geeigneteren GermanAnalyzer war deshalb nicht möglich, da zwei verschiedene Versionen Probleme hatten, die vorliegende Stoppwortliste während des Indexierungsprozesses einzubinden. Es ist davon auszugehen, dass die Leistung weiter gesteigert werden kann, indem eine leistungsfähigere Grundformenreduktion (beispielsweise Snowball[1]) eingesetzt wird.

11.1.3. Verbesserung des Übersetzungsmoduls

Während der Entwicklung hat sich gezeigt, dass das in Abschnitt 8.3 detailliert beschriebene Übersetzungsmodul keine beständige Übersetzungsleistungen bei allen Anfragen erzielt hat. Nur in 60% der Anfragen konnte ein zufriedenstellendes Ergebnis erreicht werden. Bei vier Anfragen hat die Übersetzung nicht den generellen Anforderungen genügt (Anfragen 5, 14, 18 und 23) In sechs Fällen (Anfragen 4, 6, 11, 15, 16 und 22) sind einzelne Terme mit im Thesaurus vorkommenden Phrasen übersetzt worden, bei denen der einzelne Term der Anfrage nur mit einem Term der Thesaurusphrase übereinstimmt. Dies sollte eigentlich mit der striktesten Methode (nur Übersetzung, wenn der gefundene Begriff einen Lucene-Score von 1 erhielt) der Abgleichung vermieden werden. Allerdings hatte dieser unerwartete Effekt in einigen Anfragen auch positive Eigenschaften, in dem Übersetzungen und Phrasen hinzugefügt wurden, die thematisch eng verwandt waren: So wurde als Übersetzung für den Begriff „Sicherheit“ zum Beispiel immer eine ganze Reihe von verschiedenen Sicherheitsbegriffen wie „Nationale Sicherheit“, „Militärische Sicherheit“, „Innere Sicherheit“, aber auch „Sicherheit in der Kerntechnik“ übersetzt. Einerseits war dieses Verhalten nicht geplant und streng genommen nicht erwünscht, jedoch zeigt es eine weitere Möglichkeit zur Anfragenerweiterung um potentiell relevante Terme auf. Auf die Precision-Ergebnisse der einzelnen Anfragen hat dieses Verhalten entsprechend einer Probe der anfragenorientierten

[1] vgl. auch http://snowball.tartarus.org/

Analyse (vgl. Abschnitt 10.2 auf Seite 74), wie zu erwaten war, keinen nachweisbaren Einfluss genommen. Die Anfragen mit den unerwartet erweiterten Übersetzungen verteilen sich über alle Bereich der Skala.

Im Sinne einer zuverlässigeren Übersetzung sollte jedoch bei einem weiteren Einsatz des Moduls ein phrasenorientierter Ansatz zum Zuge kommen. Ein Ansatz für ein solches System findet sich in [5].

Die Anwendung von Übersetzungen ist allerdings für den vorliegenden Auszug der Datenbasis des Fachinformationsverbundes nur für die Titel und Zusammenfassungen der Dokumente relevant: Deskriptoren und Klassifikationsangaben sind auf Deutsch verfasst. Es würde sich insgesamt empfehlen, ein separates Übersetzungsmodul zu integrieren, dass sowohl phrasenorientiert übersetzen kann und über einen erweiterten, thematisch auf internationale Politik, Recht und Wirtschaft spezialisierten Korpus verfügt.

11.2. Weitere interessante Ansätze

Neben der Korrekturen der beschriebenen Probleme ergeben sich weitere Ansätze, um die Retrievalqualität des Systems zu verbessern.

11.2.1. Automatische Verknüpfung von Termen und Phrasen

Im Rahmen der Evaluierung wurde durch einige beteiligten Fachreferenten der SWP darauf hingewiesen, dass nicht im ausreichenden Maße die einzelnen Terme in der Anfrage verknüpft werden und entsprechend viele Dokumente gefunden werden, die beispielsweise thematisch relevant, aber geopolitisch irrelevant - oder umgekehrt - sind. Voraussetzung für den automatischen Einsatz einer logischen Verknüpfung von Termen und Phrasen der reduzierten Suchanfragen ist jedoch, dass eine gewisse semantische Analyse zuvor angewendet wird, da nur auf diesem Wege der AND-Operator sinnvoll eingesetzt werden könnte: Einige Anfragen mit mehreren Themenbeschreibungen wie Anfrage 21 (Welche Probleme stellen sich bei Entwaffnung, Demobilisierung und Reintegration von Bürgerkriegs-Kombattanten?) würden sonst zu einer Anfrage umformuliert, bei der alle vier Themeninformationen verknüpft wären, obwohl auch ein Dokument relevant sein kann, dass nur ein oder zwei der relevanten Terme beinhaltet.

Eine solche semantische Analyse, ob es sich bei einem Term um eine geographische Information oder ein Thema handelt, lässt sich gegebenenfalls mit Hilfe eines Suchvorgangs des Terms die Felder *subject* und *geo* des Datenbestands oder Thesau-

rus anwenden. Ließe sich ein Term in einem Feld finden, könnte man ihn so einer Bezugsgruppe zuordnen und mit anderen Termen aus der selben Bezugsgruppe nur mit einem logischen OR verknüpfen. Bezugsgruppen untereinander würden mit einem logischen AND verknüpft. Auf diese Art und Weise könnte beispielsweise aus der Anfrage 20 (Ist Rechtsextremismus in Russland und in Polen ein Problem?) die Anfrage `Rechtsextremismus AND (Polen OR Russland)` generiert werden. Da ein solches System allerdings stark von der Nähe zum kontrollierten Vokabular abhängig ist (der Begriff des „Westlichen Balkan“ ließe sich auf diesem Wege nicht ohne weiteres als geopolitische Information entschlüsseln), wäre es nur sinnvoll einsetzbar und effektiv, wenn die Anfragen hauptsächlich aus Vokabeln des Thesaurus formuliert würden.

Eine Alternative zur einfachen Thesaurus-nahen Lösung könnte außerdem die Anwendung einer Named Entity Erkennung auf die Anfrage bieten: Ein System wie LingPipe[2], das ebenfalls im Rahmen der Entwicklung erprobt wurde, dann aber im Rahmen der Arbeit sich für einen Teilaspekt als zu aufwendig herausstellte, könnte auf sowohl das freie Vokabular der Zusammenfassungen als auch das kontrollierte Vokabular des Thesaurus trainiert werden und bei der Anwendung auf die Anfrage zurückgeben, welche Art von Bezugssystem ein Term oder eine Phrase hat.

Während sich diese Alternativen auf die Verknüpfung von Termen aus der ursprünglichen, reduzierten Anfrage bezieht, ist es im Rahmen der durch das EVM ergänzten Anfragepassagen leichter, logische Verknüpfungen zu konstruieren: Da in der Anwendung des EVM immer bekannt ist, welcher Art von Metainformation ein Term oder eine Phrase ist, lassen sich auf dieser Grundlage Terme gleicher Art einfach durch OR und verschiedene Gruppen von Termen mit AND verknüpfen.

11.2.2. Erkennung von Phrasen

Ein weiterer Ansatzpunkt zur Verbesserung des vorliegenden Systems ist die Erkennung von Phrasen. Das im Rahmen der Entwicklung des IR-Systems konzipierte Modul zur Phrasenerkennung per Thesaurus und durch die Datendateien hat sich nicht in einem akzeptablen Rahmen als leistungsfähig herausgestellt, da durch die verwendete Methodik oftmals nicht semantisch relevante Passagen der Originalanfrage als Phrase erkannt wurden. Deshalb wurde es nicht im Rahmen der Evaluierung eingesetzt.

Hier würde es sich empfehlen, ebenfalls wie bei der Analyse des Bezugssystems im vorherigen Unterkapitel, eine Named Entitiy Erkennung einzusetzen. Da mit 600.000 Dokumenten gleichviele Titel, rund 150.000 Zusammenfassungen und in einer erweiterten Ausgabe des FIV wohl auch einige zehntausende Volltexte neben den 9000

[2] vgl. auch http://www.alias-i.com/lingpipe/

Einträgen des Thesaurus zum Training eines solchen Systems zur Verfügung stünden, wären für einen solchen Einsatz die notwendigen Voraussetzungen erfüllt.

11.2.3. Weiterentwicklung von Ergänzung zu Reformulierung

Das vorliegende System ist das Ergebnis einer vergleichsweise vorsichtigen Strategie, die Leistung einer natürlichsprachlichen Anfrage durch Erweiterung und Ergänzung mit Hilfe von diversen Modulen zu verbessern. Während dies gelungen scheint, zeigt sich doch, dass professionelle Nutzer im Umgang mit der Datenbasis Anfragen formulieren, die vollkommen auf Metadaten beruhen.

Es stellt sich entsprechend die Frage, ob auch das vorliegende System in diesem Sinne „professionalisierte“ Anfragen formulieren kann, die überhaupt nicht auf der Basis der freien Felder *titel* und *abstract* aufbauen, sondern direkt auf der Ebene der Metadaten ansetzen. Dabei bleibt stetig bewusst, dass eine rein deskriptorenbasierte Anfrage den Inhalt einer natürlichsprachlichen Anfrage ggf. nicht umfassend beschreiben kann.

Mit Hilfe der in Abschnitt 8.5 beschriebenen Module müsste dabei die Anfrage, die im Rahmen des EVM zu Ermittlung des kontrollierten Vokabulars eingesetzt wird, soweit optimiert werden, dass der Prozess des EVM so zuverlässige Ergebnisse wie möglich in einer besonders kurzen Zahl von Deskriptoren extrahieren kann.

11.2.4. Nutzung von zusätzlichen Informationstypten in Anfragen

In einigen der durch die SWP vorgegebenen Anfragen kommt neben thematischen und geopolitischen Angaben auch eine temporale Komponente vor (siehe Anfrage 1, Wie haben sich die Beziehungen zwischen Nigeria und den USA in den letzten zehn Jahren entwickelt?). Da diese Art von informationellem Bezugssystem nur in 20% der vorliegenden Anfragen vorkam und die Extraktion der genauen Zeitangabe mitunter sehr schwierig ist (siehe Anfrage 18, „[...] nach Abschluss der Uruguay-Runde [...]“), wurde sie zunächst nicht in der laufenden Entwicklung des vorliegenden Systems berücksichtigt.

Da es sich bei temporalen Angaben jedoch wegen der skalaren Form um ein äußerst exaktes Mittel zur Eingrenzung der Gruppe von möglicherweise relevanten Dokumenten handelt und der temporale Bezug von zentralem Interesse für das von Prozessen und Abläufen geprägte Feld der Politikwissenschaften ist, lohnt sich gegebenenfalls die Weiterentwicklung.

Hierbei wird eine umfassende Informationsgrundlage von historischen Prozessen, Entwicklungen und Ereignissen nötig sein, um auch kompliziertere Fälle von impliziten Angaben einer temporalen Information entschlüsseln zu können.

11.2.5. Nutzung von zusätzlichen Metadatenfeldern

Während der Entwicklung des IR-Systems stellte vor allem die komplizierte XML-Struktur (vgl. Abschnitt 4.1.2 auf Seite 29) des Datenbankauszugs eine besondere Herausforderung im Vorfeld der Indexierung dar.

Es wurden im Rahmen der Indexierung bei weitem nicht alle Formen der verzeichneten Informationen auch in den Index übernommen. Beispielsweise wurden mehrere Arten von Deskriptoren zu einer Gruppe zusammengefasst und somit die hierarchischen Ansätze noch nicht ausgenutzt. Ähnlich wurde mit Klassifikationsinformationen verfahren, sodass klassifikatorische Deskriptoren nur in Form von einfachen Gruppenbezeichnungen eingesetzt werden konnten, deren hierarchische Informationen nicht berücksichtigt werden.

Es bleibt im Moment also noch ein großes Feld von Möglichkeiten für Ansätze zur besseren Gewichtung von ergänzten Metadaten - so ließe sich beispielsweise eine durch das EVM extrahierte Angabe zu einer detallierteren Untergruppe höher gewichten als die Angabe einer allgemeineren Obergruppe. Auf diesem Wege ließe sich ein umfassendes Gewichtungssystem für die unterschiedlichen hierarchischen Ebenen umsetzen.

Darüber hinaus werden derzeit im Index z.B. noch keine Autoreninformationen berücksichtigt. Da sich Autoren von (politik-)wissenschaftlichen Texten aber üblicherweise auf ein Forschungsgebiet und dementsprechend eine Region, einen Konflikt oder einen Prozess spezialisieren, kann es lohnen, eine Erweiterung des EVM auf Autorendaten zu evaluieren. Dieser Ansatz dürfte besonders für freie und offene Fragen von großem Interesse sein, da sich große Themenkomplexe meist schlecht exakt auf eine begrenzte Zahl von thematischen oder geopolitischen Deskriptoren, sondern viel eher auf ein Forschungsfeld und damit die relevanten Autoren beziehen lassen.

12. Fazit

Im Laufe dieser Arbeit wurden mehrere Ziele verfolgt und umgesetzt: Nach einem kurzen Bericht über den Kooperationspartner, die Stiftung Wissenschaft und Politik, wurde der Datenbestand des Fachinformationsverbundes sowie die unterschiedlichen Nutzungsmuster beschrieben. Im Kontext der Forschung bezüglich der Technologie der Entry Vocabulary Module wurde daraufhin ein Retrievalsystem entwickelt, dass zum besonderen Fokus hatte, die Leistungsfähigkeit eines für die Datengrundlage entwickelten, dynamischen EVM zu untersuchen. Eine Evaluierungsmethodik wurde entwickelt und daraufhin das System sowohl über fünf Evaluierungsläufe als auch über alle zur Verfügung gestellten Anfragen untersucht. Darüber hinaus wurden die einzelnen, eingesetzten Module des Systems auf Ihre Leistungsfähigkeit hin analysiert. Außerdem wurde das Retrievalsystem auf den Datenbestand GIRT3 angewendet. Eine Auswertung der Ergebnisse aus diesen Untersuchungen sowie Vorschläge zur Verbesserung der einzelnen Module schließen die Evaluierung ab. In diesem letzten Abschnitt sollen die Ergebnisse erneut thematisiert werden und in eine Empfehlung zur weiteren Entwicklung der EVM-Technologie übergehen.

12.1. Schlussfolgerungen und Erkenntnisse

Es gib mehrere zentrale Erkenntnisse aus der Erprobung des vorgestellten Systems:

Zunächst konnte die mehrsprachige Retrievalleistung durch den Einsatz der Metainformationen deutlich gesteigert werden. Dieser Erfolg ist zum Großteil dem EVM zuzuschreiben, da zuvor festgelegt wurde und hinterher belegt werden konnte, dass sowohl das Translator Modul als auch das Blind Relevance Feedback durch eine geringere Gewichtung bzw. Anzahl von Termen nur ergänzend agierten und entsprechend untergeordneten Einfluss auf die Ergebnisse genommen haben (vgl. Abschnitt 10.1.4 ab Seite 82).

Während die gesamte Retrievalqualität jedes einzelnen Laufs noch deutlich unter den Erwartungen an ein ausgereiftes System liegt, konnte die gesamte Retrievalleistung durch die zusätzlich eingesetzten Module in Recall (deutlich) und in Precision (ansatzweise) gesteigert werden.

Es hat sich allerdings gezeigt, dass auch ein Lauf der einfacheren SwpBase-Gruppe vergleichsweise gute Ergebnisse lieferte. Die Resultate der anfragenorientierten Untersuchung der Retrievalergebnisse zeigt, dass die Anfragen, die in SwpBase2 besonders erfolgreich sind, gerade in den SwpEvm-Läufen vergleichsweise schlecht abschneiden. Es lässt sich berechnen, dass die sechs erfolgreichsten (außer Anfragen 16 und 22) Anfragen in SwpBase2 durchschnittlich eine mehr als doppelt so hohe Precision hatten als der Durchschnitt dieser Anfragen über alle SwpEvm-Läufe. Daraus lässt sich, rückblickend auf das Vokabular-Problem, schlussfolgern, dass einige Anfragen anscheinend schon so günstig und entsprechend des kontrollierten Vokabulars formuliert waren, dass der Versuch, automatisch geeignetes Vokabular durch das EVM hinzuzufügen, der Qualität der Anfrage deutlich abträglich war. Ein Rückblick auf die Protokollierung des Retrievalprozess in den Outputdateien bestätigt diese Annahme (sowohl Übersetzungsmodul als auch BRF haben in diesen Anfragen zumindest nicht offensichtlich durch ihre Anfrageergänzung verschlechtert, entsprechend ist davon auszugehen, dass diese Module in diesen Anfragen keinen maßgeblichen Einfluss hatten). Es empfiehlt sich also, ein zusätzliches Modul zu entwickeln, dass im Vorfeld untersucht, ob eine Anfrage bereits über einen gewissen, nutzbaren Anteil an kontrollierten Vokabular verfügt und die entsprechenden Terme und Phrasen dann zielgerichtet auf das entsprechende Indexfeld zu richten und von einer zusätzlichen Erweiterung abzusehen. Ein solches Verfahren, das grundsätzlich die Vorteile von SwpBase2 und den SwpEvm-Läufen kombiniert, sollte in der Retrievalleistung einen deutlichen Fortschritt zeigen: Besonders bei der entsprechenden Umsetzung und Ergänzung zum Evaluierungslauf SwpEvm3 würde voraussichtlich sowohl die Retrievalleistung als auch die Homogenität der Retrievalleistung unter den Anfragen weiter deutlich gesteigert werden.

Die Ergebnisse der GIRT-Evaluierung zeigten, das die Leistungssteigerungen der eingesetzten Module auf einer Datenbasis mit umfassenden Sammlungen von Zusammenfassungen deutlich bessere Ergebnisse liefern kann als auf einer Datenbasis mit einem geringeren Anteil von Dokumenten mit Zusammenfassungen. Da im Rahmen der GIRT-Untersuchungen ebenfalls der EVM-Prozess auf fünf Elemente verkürzt wurde, liegt der Schluss nahe, dass eine verkürzte EVM-Kaskade auch für den FIV-Datensatz hilfreich gewesen sein könnte. Nachträgliche Untersuchungen dieser These haben jedoch das Gegenteil belegt: Sowohl Precision als auch Recall nahmen bei der Anwendung auf den FIV-Bestand mit jedem weiteren aktivierten Element weiter zu. Dies führt eher zu der Annahme, dass (voraussichtlich bis zu einem bestimmten Punkt) komplexere Muster zu Extraktion von Metadaten bei einem Datenbestand wie dem des FIV die Retrievalqualität steigern können.

Gleichzeitig ist zu beachten, dass im Vergleich zu einer professionellen Suchanfrage

(vgl. Abschnitt 4.2.1 ab Seite 33) nur eine deutlich geringere Anzahl von Dokumenten gefunden werden konnte. Es ist zu vermuten, dass beispielsweise durch einige sehr allgemeine Klassifikationsangaben die Suchergebnisse nicht präziser sondern eher unschärfer wurden. Es würde sich also insgesamt empfehlen, die hierarchisch detailliertesten, präzisesten, verfügbaren Einträge des Thesaurus (soweit verfügbar) und der Klassifikation zu verwenden. Diese Empfehlung ist allerdings erneut im Kontext des Konkretisierungsgrades der Anfrage zu verstehen - eine Anfrage mit konkreten Themenangaben (vgl. Anfrage 9) mag auf diese Weise profitieren, eine weit gefasste Anfrage (vgl. Anfrage 11) könnte von zu spezifischen Deskriptoren vom Thema abgelenkt werden.

Eine Schlüsselfrage zur Retrievalleistung von Feedback-Systemen ist die der Qualität der ursprünglichen Anfrage. Nur eine geeignete, ursprüngliche Anfrage kann zu einer Ergebnisauflistung führen, bei der die am höchsten bewerteten Dokumente zu einem großen Anteil relevant sind. Entsprechend sollte neben der möglichen Weiterentwicklung eines dynamischen EVM gleichzeitig an der Vorbereitung der ursprünglichen Anfrage für den Feedback-Lauf gearbeitet werden. In diesem Rahmen können beispielsweise die Terme der Anfragen auf Phrasen hin untersucht werden oder, wie weiter oben beschrieben, bereits vorhandenes, kontrolliertes Vokabular auch konkret auf das entsprechend passende Indexfeld gerichtet werden. Durch die Qualität der ursprünglichen Anfrage steigt die Qualität der gefundenen, untersuchten Dokumente, die Auswertung der Deskriptoren führt potentiell zu weniger, höher bewerteten und auch geeigneteren Deskriptoren. Weiterhin ließe sich ein iteratives System implementieren, dass auf den bereits verbesserten Ergebnissen eines vorhergehenden SwpEvm-Laufs beruht.

Bei der Evaluierung des vorliegenden, automatischen Systems konnte darüber hinaus belegt werden, dass durch den Einsatz der beschriebenen Module eine größere Anzahl von Anfragen überdurchschnittliche Retrievalergebnisse erzielen konnten. Im Bezug auf die eingesetzten Parameter hat sich herausgestellt, dass die Auswertung einer geringeren Anzahl von Dokumenten in Kombination mit einem niedrigen Score-Schwellenwert zur Erweiterung eine vergleichsweise geringe Anzahl von Anfragen deutlich in der Retrievalqualität fördert. Im Gegensatz dazu erzielte die Kombination einer großen Anzahl von untersuchten Dokumenten und einem hohen Schwellenwert eine größere Anzahl von Anfragen mit überdurchschnittliche Precision-Ergebnissen. In weiteren Untersuchungen könnte beispielsweise die Parameterkombination einer niedrigen Anzahl von untersuchten Dokumenten mit einem hohen Schwellenwert sowie viele weitere Umsetzungen, bspw. die Anpassung der untersuchten Dokumenten an einen Anteil aller gefundenen Dokumente (Top 5%) oder der Einsatz von Schwellenwerten zur Auswahl der berücksichtigten Dokumente, analysiert werden.

Eine Realisierung eines Entry Vocabulary Moduls ist grundsätzlich zu empfehlen, da es sowohl für professionelle als auch semiprofessionelle Nutzer einen Vorteil in der Retrievalqualität liefern kann.

Für natürlichsprachliche Anfragen - also wahrscheinlich Anfragen der Gruppe der semiprofessionellen Nutzer wie beispielsweise Nutzer eines Online-Portals, das die Datenbasis des FIV zugänglich macht - kann sowohl die interaktive Umsetzung als auch der automatische Einsatz eines EVM Vorteile bringen. Allerdings sind dabei die diversen, im vorhergehenden Unterkapitel genannten Punkte zu berücksichtigen, die die Ergebnisse eines solchen automatischen EVM über die in der Evaluierung aufgezeigten Leistungen hinweg verbessern können. Ein Entry Vocabulary Modul kann helfen, diejenigen Anfragen, die nicht im bereits über einen gewissen Anteil an kontrolliertem Vokabular verfügen, in ihrer Retrievalqualität zu steigern.

Für der professionellen Einsatz kann ein EVM ebenfalls nützlich sein. In diesem Umfeld würde sich allerdings empfehlen, ein interaktives System zu implementieren, mit dem den professionellen Nutzern das manuelle Suchen von geeigneten Deskriptoren abgenommen wird. Es ist zu erwarten, dass eine mit professionellem Verständnis ausgewählte Gruppe von Deskriptoren die Leistungsfähigkeit eines automatischen Systems deutlich übersteigt.

Während die Entwicklung eines ersten Prototyps für ein Entry Vocabulary Moduls geglückt scheint, kann dies nur der allererste Schritt gewesen sein, ein solches System zu entwickeln und für den Einsatz vorzubereiten. Das vorgestellte Konzept ist sehr variable, daher könnte durch weitere Testläufe mit verschiedenen Parameterkombinationen, Kaskadenlängen und Möglichkeiten zur Optimierung der ursprünglichen Anfrage festgestellt werden, wie sich das System weiter optimieren lässt. Im Rahmen eines einzelnen Evaluierungsprozesses war die Erforschung der Parameter nur im Ansatz möglich. Es zeigten sich jedoch bereits in diesem prototypischen Stadium positive Effekte durch den Einsatz eines Entry Vocabulary Moduls, sodass eine Weiterentwicklung des Systems zu empfehlen ist.

Teil VI.

Anhang

Evaluierungsanfragen

1. Wie haben sich die Beziehungen zwischen Nigeria und den USA in den letzten zehn Jahren entwickelt?
2. Wie ist der aktuelle Stand des Wiederaufbaus Afghanistans?
3. Welche Faktoren bestimmen die Beziehungen zwischen China und der EU / den einzelnen EU-Ländern?
4. Welche Gefährdungen bestehen für die maritime Sicherheit in Südostasien?
5. Welche Rolle spielen Religion und religiöse Faktoren in den internationalen Beziehungen?
6. Welches sind die Bestimmungsfaktoren und Probleme der sicherheitspolitischen Beziehungen der USA zum asiatisch-pazifischen Raum?
7. Welche Rolle spielen Massenvernichtungswaffen und Terrorismusbekämpfung in den transatlantischen Beziehungen?
8. Wie gestaltet sich die zivil-militärische Zusammenarbeit beim Wiederaufbau von Nachkriegsgesellschaften?
9. Welche Erdölpolitik betreibt Libyen?
10. Wie verhält sich China gegenüber dem nordkoreanischen Atomkonflikt?
11. Wie gestaltet die erweiterte Europäische Union ihre Beziehungen gegenüber den europäischen Staaten der ehemaligen Sowjetunion und gegenüber dem Westlichen Balkan?
12. Welche Minderheitenpolitik betreiben die EU-Staaten?
13. Welches sind die Probleme und Perspektiven einer demokratischen Neuordnung im Nahen und Mittleren Osten?

14. Welche Unterschiede und Gemeinsamkeiten bestehen zwischen der europäischen/deutschen und der amerikanischen Lateinamerika-Politik?

15. Welche Haltung nehmen die britischen Parteien gegenüber der EU ein?

16. Welches sind die zentralen Herausforderungen für die Europäische Sicherheits- und Verteidigungspolitik?

17. Welche Rolle spielt die EU als globaler Akteur im Politikfeld Handelspolitik?

18. Wie hat sich der WTO-Prozess nach Abschluss der Uruguay-Runde weiterentwickelt?

19. Welche Bedeutung hat Wettbewerbspolitik als Thema für die WTO?

20. Ist Rechtsextremismus in Russland und in Polen ein Problem?

21. Welche Probleme stellen sich bei Entwaffnung, Demobilisierung und Reintegration von Bürgerkriegs-Kombattanten?

22. Wie hat sich die Gemeinsame Außen- und Sicherheitspolitik der EU in den letzten Jahren entwickelt?

23. Wie gestaltet sich die Zukunft des mediterranen Raums aus europäischer Sicht?

24. Welches sind die zentralen Probleme bei Wiederaufbau und Friedenssicherung in Nachkriegs-Gesellschaften?

25. Wie werden Globalisierungsprozesse in der Arabischen Welt perzipiert und welche Auswirkungen haben sie dort?

Anfrageanalyse

Die im Vorfeld der Evaluierung festgelegten Suchanfragen werden im Folgenden auf Ansatzpunkte hin untersucht.

1. Wie haben sich die Beziehungen zwischen Nigeria und den USA in den letzten zehn Jahren entwickelt?

   Geopolitischer Bezug: Nigeria, USA

   Thematischer Bezug: Beziehungen

   Temporaler Bezug: in den letzten zehn Jahren

   Kommentar: Die Frage hat einen klar bilateralen geopolitischen Bezug, bleibt aber im thematischen Bezug sehr offen.

2. Wie ist der aktuelle Stand des Wiederaufbaus Afghanistans?

   Geopolitischer Bezug: Afghanistan

   Thematischer Bezug: Wiederaufbau

   Temporaler Bezug: aktuelle Stand

   Kommentar: Die Frage hat im Bezug auf Geoinformationen und Thematik einen klaren und geschlossenen, unilateralen Charakter.

3. Welche Faktoren bestimmen die Beziehungen zwischen China und der EU / den einzelnen EU-Ländern?

   Geopolitischer Bezug: China, EU, einzelnen EU-Ländern

   Thematischer Bezug: Faktoren, Beziehungen

   Temporaler Bezug: —

   Kommentar: Der offene thematische Bezug „Beziehungen“ macht die Frage unspezifisch, gleichzeitig ist die Staffelung der Geoinformationen ungünstig in Relation zu setzen. Darüber hinaus lässt sich „den einzelnen EU-Ländern“ nicht ohne ein zusätzliches Modul auf die Namen der einzelnen EU-Länder auflösen (nach dieser Frage ist schließlich auch jedes Dokument relevant, bei dem die Beziehungen nur eines EU-Landes zu China beschrieben wird).

4. Welche Gefährdungen bestehen für die maritime Sicherheit in Südostasien?

   Geopolitischer Bezug: Südostasien

   Thematischer Bezug: Gefährdungen, maritime Sicherheit

   Temporaler Bezug: —

Kommentar: Die Frage hat einen geopolitischen Bezug, der sich auf eine Region bezieht. Die Themensetzung ist durch den vagen Begriff „Gefährdungen“ vergleichsweise offen.

5. Welche Rolle spielen Religion und religiöse Faktoren in den internationalen Beziehungen?

   Geopolitischer Bezug: —

   Thematischer Bezug: Religion, religiöse Faktoren, internationale Beziehungen

   Temporaler Bezug: —

   Kommentar: Die Anfrage ist sehr offen formuliert, da sie keinen geopolitischen Bezug enthält. Die Themen „Religion“ und „religiöse Faktoren“ geben zwar eine Richtung vor, sind aber nicht sehr spezifisch.

6. Welches sind die Bestimmungsfaktoren und Probleme der sicherheitspolitischen Beziehungen der USA zum asiatisch-pazifischen Raum?

   Geopolitischer Bezug: USA, asiatisch-pazifischer Raum

   Thematischer Bezug: sicherheitspolitische Beziehungen

   Temporaler Bezug: —

   Kommentar: Die Frage hat einen bilateralen geopolitischen Bezug zwischen den USA und der Region bzw. den Staaten des asiatisch-pazifischen Raumes, der sich nicht eindeutig auflösen lässt. Sie bietet einen recht offenen thematischer Bezug.

7. Welche Rolle spielen Massenvernichtungswaffen und Terrorismusbekämpfung in den transatlantischen Beziehungen?

   Geopolitischer Bezug: transatlantischen

   Thematischer Bezug: Massenvernichtungswaffen, Terrorismusbekämpfung

   Temporaler Bezug: —

   Kommentar: Die Frage hat einen regionalen, geopolitischen Bezug sowie zwei klare Themenangaben.

8. Wie gestaltet sich die zivil-militärische Zusammenarbeit beim Wiederaufbau von Nachkriegsgesellschaften?

   Geopolitischer Bezug: —

   Thematischer Bezug: zivil-militärische Zusammenarbeit, Wiederaufbau, Nachkriegsgesellschaften

   Temporaler Bezug: —

Kommentar: Die Frage enthält keinerlei geopolitischen Bezug, bietet aber mehrere eindeutige thematische Anhaltspunkte.

9. Welche Erdölpolitik betreibt Libyen?

   Geopolitischer Bezug: Libyen

   Thematischer Bezug: Erdölpolitik

   Temporaler Bezug: —

   Kommentar: Die Frage ist eindeutig und enthält klare Bezüge.

10. Wie verhält sich China gegenüber dem nordkoreanischen Atomkonflikt?

    Geopolitischer Bezug: China, nordkoreanischen

    Thematischer Bezug: Atomkonflikt

    Temporaler Bezug: —

    Kommentar: Die Frage hat einen klar bilateralen geopolitischen Bezug und erwähnt ein eindeutiges Thema.

11. Wie gestaltet die erweiterte Europäische Union ihre Beziehungen gegenüber den europäischen Staaten der ehemaligen Sowjetunion und gegenüber dem Westlichen Balkan?

    Geopolitischer Bezug: erweiterte Europäische Union, europäische Staaten, ehemaligen Sowjetunion, Westlichen Balkan

    Thematischer Bezug: Beziehungen

    Temporaler Bezug: ehemaligen Sowjetunion

    Kommentar: Die Frage hat einen klar multilateralen Charakter und mehrere geopolitische Bezüge, der thematische Bezug ist sehr offen.

12. Welche Minderheitenpolitik betreiben die EU-Staaten?

    Geopolitischer Bezug: EU-Staaten

    Thematischer Bezug: Minderheitenpolitik

    Temporaler Bezug: —

    Kommentar: Die Frage hat grundsätzlich sehr klare geopolitische und thematische Informationen, ist aber inhaltlich mehrdeutig „EU-Staaten“ (Staaten im Rahmen der EU, Politik der einzelnen EU-Staaten).

13. Welches sind die Probleme und Perspektiven einer demokratischen Neuordnung im Nahen und Mittleren Osten?

    Geopolitischer Bezug: Nahen, Mittleren Osten

Thematischer Bezug: Probleme, Perspektiven, demokratische Neuordnung

Temporaler Bezug: —

Kommentar: Die Frage gibt einen Richtung vor, um das Thema einzugrenzen. Die Geoinformationen sind nützlich, beziehen sich allerdings auf im Vergleich zu Ländernamen deskriptiv unschärfere Regionen.

14. Welche Unterschiede und Gemeinsamkeiten bestehen zwischen der europäischen/deutschen und der amerikanischen Lateinamerika-Politik?

Geopolitischer Bezug: europäischen, deutschen, amerikanischen, Lateinamerika

Thematischer Bezug: Unterschiede, Gemeinsamkeiten, Lateinamerika-Politik

Temporaler Bezug: —

Kommentar: Der geopolitischen Bezug lässt sich feststellen, allerdings ist die Verknüpfung von Europa, Deutschland, Amerika und Lateinamerika inhaltlich nicht abzubilden. Thematisch ist diese Anfrage nichteindeutig.

15. Welche Haltung nehmen die britischen Parteien gegenüber der EU ein?

Geopolitischer Bezug: britischen, EU

Thematischer Bezug: Haltung, britischen Parteien, EU

Temporaler Bezug: —

Kommentar: Klare geopolitische Bezüge, die Angabe „Haltung“ ist aber sehr offen.

16. Welches sind die zentralen Herausforderungen für die Europäische Sicherheits- und Verteidigungspolitik?

Geopolitischer Bezug: Europäische

Thematischer Bezug: Herausforderungen, Sicherheits- und Verteidigungspolitik

Temporaler Bezug: —

Kommentar: Die Frage hat einen klar unilateralen, geopolitischen Bezug, bleibt aber im thematischen Bezug sehr offen.

17. Welche Rolle spielt die EU als globaler Akteur im Politikfeld Handelspolitik?

Geopolitischer Bezug: EU, globaler

Thematischer Bezug: Handelspolitik

Temporaler Bezug: —

Kommentar: Die Frage hat einen klar unilateralen, geopolitischen Bezug, bleibt aber im thematischen Bezug sehr offen.

18. Wie hat sich der WTO-Prozess nach Abschluss der Uruguay-Runde weiterentwickelt?

    Geopolitischer Bezug: —

    Thematischer Bezug: WTO, Uruguay-Runde

    Temporaler Bezug: nach Abschluss der Uruguay-Runde

    Kommentar: Der geopolitische Bezug ist gänzlich offen, die vermeintliche Geoinformation „Uruguay“ ist Teil des thematischen Bezugs. Das Schlagwort WTO lässt sich erfolgreich bei der Suche einsetzen.

19. Welche Bedeutung hat Wettbewerbspolitik als Thema für die WTO?

    Geopolitischer Bezug: —

    Thematischer Bezug: Wettbewerbspolitik, WTO

    Temporaler Bezug: —

    Kommentar: Keinerlei geopolitischer Bezug, der thematische Bezug zu „Wettbewerbspolitik“ und „WTO“ ist jedoch eindeutig.

20. Ist Rechtsextremismus in Russland und in Polen ein Problem?

    Geopolitischer Bezug: Russland, Polen

    Thematischer Bezug: Rechtsextremismus

    Temporaler Bezug: —

    Kommentar: Die Frage bietet klare geopolitische Informationen und auch ein thematisches Schlagwort, sie bezieht sich aber deutlich auf zwei verschiedene Antwortkomplexe (Bezug Russland vs. Bezug Polen).

21. Welche Probleme stellen sich bei Entwaffnung, Demobilisierung und Reintegration von Bürgerkriegs-Kombattanten?

    Geopolitischer Bezug: —

    Thematischer Bezug: Probleme, Entwaffnung, Demobilisierung, Reintegration, Bürgerkriegs-Kombattanten

    Temporaler Bezug: —

    Kommentar: Der thematische Bezug ist in vielen Schlagwörtern ausgedrückt, es wurden keine geopolitischen Informationen angegeben.

22. Wie hat sich die Gemeinsame Außen- und Sicherheitspolitik der EU in den letzten Jahren entwickelt?

    Geopolitischer Bezug: EU

Thematischer Bezug: Gemeinsame Außen- und Sicherheitspolitik

Temporaler Bezug: in den letzten Jahren

Kommentar: Die Bezeichnung „Gemeinsame Außen- und Sicherheitspolitik“ ist in Kombination mit dem Geopolitischer Bezug „EU“ eindeutig.

23. Wie gestaltet sich die Zukunft des mediterranen Raums aus europäischer Sicht?

Geopolitischer Bezug: mediterranen Raums, europäischer

Thematischer Bezug: Zukunft mediterranen Raums

Temporaler Bezug: Zukunft

Kommentar: Die Information zum geopolitischen Bezug durch „mediterranen Raums“ ist recht unscharf, ebenfalls bietet der thematische bzw. temporale Bezug „Zukunft“ keine eindeutigen Anhaltspunkte.

24. Welches sind die zentralen Probleme bei Wiederaufbau und Friedenssicherung in Nachkriegs-Gesellschaften?

Geopolitischer Bezug: —

Thematischer Bezug: Probleme, Wiederaufbau, Friedenssicherung, Nachkriegs-Gesellschaften

Temporaler Bezug: —

Kommentar: Die Frage umfasst mehrere eindeutige, themenbezogene Schlagwörter und keinen geopolitischen Bezug.

25. Wie werden Globalisierungsprozesse in der Arabischen Welt perzipiert und welche Auswirkungen haben sie dort?

Geopolitischer Bezug: Arabischen Welt

Thematischer Bezug: Globalisierungsprozesse, Auswirkungen

Temporaler Bezug: —

Kommentar: Der geopolitsche Bezug beschreibt eine Region und ist damit recht vage, der thematische Bezug hat Oberbegriffscharakter.

Literaturverzeichnis

[1] Airio, Eija: Word normalization and decompounding in mono- and bilingual IR. In: *Information Retrieval* Band 9 (2006), Juni, Nr. Ausgabe 3

[2] Arntz, Reiner ; Picht, Heribert ; Mayer, Felix: *Einführung in die Terminologiearbeit.* Fünfte Ausgabe. Hildesheim : Olms, 2004

[3] Attar, Rony ; Fraenkel, Aviezri: Local Feedback in Full-Text Retrieval Systems. In: *Journal of the Association for Computing Machinery* 24 (1977), July, Nr. 3, S. 397–417

[4] Baeza-Yates, Ricardo ; Ribeiro-Neto, Berthier: *Modern Information Retrieval.* New York, USA : Addison Wesley, 1999

[5] Ballesteros, Lisa ; Croft, Bruce: Phrasal Translation and Query Expansion Techniques for Cross-language Information Retrieval. In: *Proceedings of the 20th ACM SIGIR Conference on Research and Development in Information Retrieval.* Philadelphia, USA : ACM Press, 1997, S. 84–91

[6] Barth, Viola ; Pfister, Joachim: *Programmdokumentation der Projektgruppe Trec_eval.* Universität Hildesheim, 2003

[7] Bauermeister, Matthias. *Fragestellung 20: Ist Rechtsextremismus in Rußland ein Problem?* Interne Evaluierungsauswertung. Juli 2006

[8] Braschler, Martin ; Ripplinger, Bärbel: Stemming and Decompounding for German Text Retrieval. In: *Advances in Information Retrieval: 25th European Conference on IR Research, ECIR 2003, Pisa, Italy, April 14-16, 2003. Proceedings*, Springer, April 2003, S. 177–192

[9] Buckland, Michael ; Chen, Aitao ; Chen, Hui-Min ; Kim, Youngin ; Lam, Byron ; Larson, Ray ; Norgard, Barbara ; Purat, Jacek ; Gey, Fredric: Mapping Entry Vocabulary to Unfamiliar Metadata Vocabularies. In: *D-Lib Magazine* 5 (1999), January 1999, Nr. 1

[10] Crestani, Fabio ; Lalmas, Mounia ; Rijsbergen, Cornelis J. V. ; Campbell, Iain: Is This Document Relevant? . . . Probably: A Survey of Probabilistic Models in Information Retrieval. In: *ACM Computing Surveys* 30 (1998), Dezember, Nr. 4, S. 528–552

[11] Dunning, Ted: Accurate Methods for the Statistics of Surprise and Coincidence. In: *Computational Linguistics* Band 19 (1994), Nr. Nummer 1, S. Seiten 61–74

[12] Fachinformationsverbund IBLK: *Feldgruppenstruktur des deutschsprachigen Euro-Thesaurus*. Stiftung Wissenschaft und Politik, 2004, S. 207–216. – Verfügbar unter http://www.fiv-iblk.de/ip/dokumente/fgs.pdf, verifiziert am 19. September 2006

[13] Fachinformationsverbund IBLK: *FIV-Regionalklassifikation*. Website. Juli 2006. – Verfügbar unter http://www.fiv-iblk.de/ip/dokumente/geoklass_d.pdf, verifiziert am 19. September 2006

[14] Fachinformationsverbund IBLK: *FIV-Sachklassifikation*. Website. Juli 2006. – Verfügbar unter http://www.fiv-iblk.de/ip/dokumente/sachklass_d.pdf, verifiziert am 19. September 2006

[15] Ferber, Reginald: *Information Retrieval*. Bd. 1. 1. dpunkt.verlag, Heidelberg, 2003

[16] Fuhr, Norbert: *Allgemeine Informationen über die Fachgruppe*. Website. Januar 1996. – Verfügbar unter http://www.uni-hildesheim.de/fgir/index.php?option=com_contenttask=viewid=14Itemid=41, verifiziert am 30. September 2006

[17] Gey, Fredric ; Buckland, Michael ; Chen, Aitao ; Larson, Ray: Entry vocabulary - a technology to enhance digital search. In: *Proceedings of the first international conference on Human language technology research* (2001)

[18] Gey, Fredric ; Chen, Aitao: Phrase Discovery for English and Cross-language Retrieval at TREC 6. In: *Text REtrieval Conference*, 1997, S. 637–647

[19] Gey, Fredric ; Chen, Hui-Min ; Norgard, Barbara ; Kim, Youngin ; Buckland, Michael ; Chen, Aitao ; Larson, Ray ; Lam, Byron ; Purat, Jacek: Advanced Search Technologies for Unfamiliar Metadata / University of California. 1999. – Forschungsbericht

[20] Gey, Fredric ; Jiang, Hailing: English-German Cross-Language Retrieval for the GIRT Collection - Exploiting a Multilingual Thesaurus. In: *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, 1999

[21] Gospodnetic, Otis: *Parsing, indexing, and searching XML with Digester and Lucene*. Website. Juni 2003. – Verfügbar unter http://www-128.ibm.com/developerworks/java/library/j-lucene/, verifiziert am 19. September 2006

[22] Gospodnetic, Otis ; Hatcher, Erik: *Lucene in Action*. Manning, 2005

[23] Grossmann, David A. ; Frieder, Ophir ; Croft, Bruce (Hrsg.): *Information Retrieval: Algorithms and Heuristics*. Zweite Ausgabe. Springer, 2004 (The Kluwer International Series on Information Retrieval)

[24] Hackl, René ; Mandl, Thomas ; Womser-Hacker, Christa: Mono- and Crosslingual Retrieval Experiments at the University of Hildesheim. In: Peters, Carol (Hrsg.) ; Clough, Paul (Hrsg.) ; Gonzalo, Julio (Hrsg.) ; Kluck, Michael (Hrsg.) ; Jones, Gareth (Hrsg.) ; Magnini, Bernard (Hrsg.): *Multilingual Information Access for Text, Speech and Images: Results of the fifth CLEF Evaluation Campaign* Universität Hildesheim, 2005, S. 165–169

[25] Hartmann, Reinhard ; James, Gregory: *Dictionary of lexicography.* Erste Ausgabe. Routledge, London, 1998

[26] Hartmann, Ulrich ; Steinmeier, Frank-Walter ; Vogel, Heinrich ; Bertram, Christoph. *Reden zum 40jährigen Bestehen, 1.Juli 2002.* Verfügbar unter http://www.swp-berlin.org/common/get_document.php?id=1034, verifiziert am 19. September 2006. Juli 2002

[27] Hellweg, Heiko ; Krause, Jürgen ; Mandl, Thomas ; Marx, Jutta ; Müller, Matthias ; Mutschke, Peter ; Strötgen, Robert: Treatment of Semantic Heterogeneity in Information Retrieval / IZ Sozialwissenschaften, Bonn. 2001 (23). – Forschungsbericht

[28] Kluck, Michael: Die GIRT-Testdatenbank als Gegenstand informationswissenschaftlicher Evaluation. Konstanz, Deutschland : UVK Verlagsgesellschaft, Oktober 2004, S. 247–268

[29] Kuhlen, Rainer: *Informationsethik.* UVK Verlagsgesellschaft, 2004

[30] Lima, Erika de ; Pedersen, Jan: Phrase recognition and expansion for short, precision-biased queries based on a query log. In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* ACM, ACM Press, 1999, S. 145 – 152

[31] Lyman, Peter ; Varian, Hal: *How Much Information? 2003.* Website. 2003. – Verfügbar unter http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/, verifiziert am 19. September 2006

[32] Magennis, Mark ; Rijsbergen, Cornelis J.: The potential and actual effectiveness of interactive query expansion. In: *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, 1997, S. 324 – 332

[33] Mair, Stefan ; Paul, Michael ; Schneckener, Ulrich: Wissenschaftliche Politikberatung am Beispiel der Stiftung Wissenschaft und Politik. In: Bröchler, Stephan (Hrsg.) ; Schützeichel, Rainer (Hrsg.): *Grundwissen Politikberatung - Ein Handbuch.* Stuttgart, 2005. – Vorab verfügbar unter http://www.swp-berlin.org/common/get_document.php?id=1450, verifiziert am 19. September 2006

[34] Norgard, Barbara: Entry Vocabulary Modules and Agents / University of Berkeley. 1998. – Forschungsbericht

[35] Online Computer Library Center: *Introduction to Dewey Decimal Classification.* Online Nachdruck (Auszug). 2003. – Verfügbar unter http://www.oclc.org/dewey/versions/ddc22print/intro.pdf, verifiziert am 19. September 2006

[36] Petras, Vivien: GIRT and the Use of Subject Metadata for Retrieval. In: *Multilingual Information Access for Text, Speech and Images* Band 3491/2005 (2005), S. 298–309

[37] Petras, Vivien: How One Word Can Make all the Difference - Using Subject Metadata for Automatic Query Expansion and Reformulation / Cross Language Evaluation Forum. Wien, Österreich, September 2005. – Forschungsbericht

[38] Petras, Vivien ; Perleman, Natalia ; Gey, Fredric: Using Thesauri in Cross-Language Retrieval of German and French Indexed Collections. In: *Advances in Cross-Language Information Retrieval*, 2003, S. 349–362

[39] Rijsbergen, Keith van ; Information Retrieval Group, University of G. (Hrsg.): *Information Retrieval.* Bd. 1. Zweite Ausgabe. Butterworths, London, 1979

[40] Schatz, Bruce ; Johnson, Eric ; Cochrane, Pauline: Interactive Term Suggestion for Users of Digital Libraries: Using Subject Thesauri and Co-occurence Lists for Information Retrieval. In: *First ACM International Conference on Digital Libraries Proceedings.* Bethesda, MD, USA : ACM, 1996, S. 126–133

[41] Smith, Alastair: Search features of digital libraries. In: *Information Research* 5 (2000), April, Nr. 3

[42] Stiftung Wissenschaft und Politik: *Stufen der Entwicklung des Fachinformationsverbunds Internationale Beziehungen und Länderkunde.* Online-PDF. – Verfügbar unter http://www.fiv-iblk.de/fiv/dokumente/fiv_geschichte.pdf, verifiziert am 19. September 206

[43] Stiftung Wissenschaft und Politik: *Erschließung der Datenbasis.* Website. 2005. – Verfügbar unter http://fiv-iblk.de/db/erschliessung.htm, verifiziert am 19. September 2006

[44] Stiftung Wissenschaft und Politik: *Der Fachinformationsverbund Internationale Beziehungen und Länderkunde.* Website. 2005. – Verfügbar unter http://www.fiv-iblk.de/fiv/information.htm, verifiziert am 19. September 2006

[45] Stiftung Wissenschaft und Politik: *Inhalte der Datenbasis.* Website. 2005. – Verfügbar unter http://www.fiv-iblk.de/db/inhalte.htm, verifiziert am 19. September 2006

[46] Stiftung Wissenschaft und Politik: *Aufgaben des Fachinformationsbereichs*. Website. 2006. – Verfügbar unter http://www.swp-berlin.org/other/struktur.php?page=3, verifiziert am 19. September 2006

[47] Stiftung Wissenschaft und Politik: *Organisationsdiagramm des Deutschen Instituts für Internationale Politik und Sicherheit der SWP*. Website. 2006. – Verfügbar unter http://swp-berlin.org/other/struktur.php?page=5, verifiziert am 19. September 2006

[48] Kap. Appendix 1. In: TREC: *Common Evaluation Measures*. Department of Commerce, National Institute of Standards and Technology, November 2005. – Verfügbar unter http://trec.nist.gov/pubs/trec14/appendices/CE.MEASURES05.pdf, verifiziert am 19. September 2006

[49] Voorhees, Ellen M.: Query Expansion using Lexical-Semantic Relations. In: *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1994, S. 61–69

[50] Xu, Jinxi ; Croft, Bruce: Query Expansion Using Local and Global Document Analysis. In: *Proceedings of the Nineteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1996, S. 4–11

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig verfasst habe und in diesem Prozess ausschließlich Hilfsmittel und Quellen verwendet wurden, die ich als solche kenntlich gemacht habe.